Adaptive time-stepping for incompressible flow part I: scalar advection-diffusion

Gresho, Philip M.; Silvester, David J.; Griffiths, David

Published in:
SIAM Journal on Scientific Computing

DOI:
10.1137/070688018

Publication date:
2008

SIAM J. SCI. COMPUT.
Vol. 30, No. 4, pp. 2018–2054

ADAPTIVE TIME-STEPPING FOR INCOMPRESSIBLE FLOW PART I: SCALAR ADVECTION-DIFFUSION*

PHILIP M. GRESHO†, DAVID F. GRIFFITHS‡, AND DAVID J. SILVESTER§

Abstract. Even the simplest advection-diffusion problems can exhibit multiple time scales. This means that robust variable step time integrators are a prerequisite if such problems are to be efficiently solved computationally. The performance of the second order trapezoid rule using an explicit Adams–Bashforth method for error control is assessed in this work. This combination is particularly well suited to long time integration of advection-dominated problems. Herein it is shown that a stabilized implementation of the trapezoid rule leads to a very effective integrator in other situations: specifically diffusion problems with rough initial data; and general advection-diffusion problems with different physical time scales governing the system evolution.

Key words. time-stepping, adaptivity, convection-diffusion

AMS subject classifications. 65M12, 65M15, 65M20

DOI. 10.1137/070688018

1. Background and context. The adaptive time-stepping algorithm that is the focus of this work is certainly not new. We consider the simplest Adams–Bashforth–Moulton pair. A version of our algorithm is hard-wired as the MATLAB function `ode23t`, see [24], and the underlying methodology is discussed in any many textbooks on the numerical solution of ODEs. See, for example, Henrici [13, p. 258] where estimation of the truncation error is discussed, or Iserles [16, p. 78], where step-doubling and halving is described.

The aim of this work is to assess the performance of this integrator in the context of method-of-lines discretization of PDEs that arise in incompressible flow modelling. In particular, we hope to provide insight into the role of adaptive time-stepping in the context of modelling multiple physical timescales. For this purpose it suffices to restrict our attention to the following simple model of scalar advection-diffusion:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0 \quad \text{on} \quad 0 \le x \le 1, \tag{1.1}$$

together with the initial condition $u(x,0) = u_0(x)$, and boundary conditions (BCs)

$$u(0,t) = u_L \quad \text{and either,} \tag{1.2}$$

$$u(1,t) = u_R \quad \text{or} \quad \frac{\partial u}{\partial x}(1,t) = 0, \tag{1.3}$$

where $a \ge 0$ (the advecting velocity), $\nu \ge 0$ (diffusivity), and u_L and u_R are given constants. In part II, we build on the foundation laid in this paper and consider the potential of the integrator in the context of solving the Navier–Stokes equations.

*Received by the editors April 11, 2007; accepted for publication (in revised form) September 7, 2007; published electronically May 14, 2008. This collaboration was supported by EPSRC grant GR/R26092/1.

http://www.siam.org/journals/sisc/30-4/68801.html

†Livermore, CA, USA (pgresho@comcast.net).

‡Mathematics Division, University of Dundee, DD1 4HN, Scotland, UK (dfg@maths.dundee.ac.uk).

§School of Mathematics, University of Manchester, M13 9PL, UK (d.silvester@manchester.ac.uk).

Spatial discretization will, throughout this paper, be carried out using the standard Galerkin approximation with piecewise linear finite elements on an N-element mesh. This leads to the system of coupled ODEs

$$M\dot{\mathbf{u}} + A\mathbf{u} = \mathbf{f}; \quad \mathbf{u}(0) = \mathbf{u}_0, \tag{1.4}$$

where the vector $\mathbf{f}$ arises from the BC's (and is zero in the homogeneous case) and, for a Dirichlet BC at $x = 1$, $\mathbf{u}(t) := (U_1(t), U_2(t), \ldots U_{N-1}(t))^T$, where $\{U_j\}$ are the nodal values of the finite element approximation. With a Neumann BC at $x = 1$ the vector $\mathbf{u}(t)$ will contain N components. Thus, for a uniform subdivision of intervals of length $h = 1/N$, the jth component of (1.4) is the second order centered finite difference equation

$$\frac{1}{6}\left[\dot{U}_{j-1} + 4\dot{U}_j + \dot{U}_{j+1}\right] + \frac{a}{2h}\left[U_{j+1} - U_{j-1}\right] - \frac{\nu}{h^2}\left[U_{j-1} - 2U_j + U_{j+1}\right] = 0. \tag{1.5}$$

For further details, see, e.g., Gresho and Sani [7, p. 40]. The matrix A in (1.4) is the sum of a symmetric positive-definite diffusion matrix K and a skew symmetric convection matrix C, so as to properly mimic their continuous (operator) counterparts. M is the mass matrix associated with a discrete L_2 projection operator.

The adaptive time-stepping algorithm that is applied to the ODE system (1.4) is a refined version of the "smart integrator" advocated by Gresho and Sani [7, section 2.7.3–4]. Our algorithm has three ingredients: time integration, the time step selection method, and stabilization of the integrator. We discuss each of these separately below.

Time integration. According to the trapezoid rule (TR), given a vector $\mathbf{u}_n \approx \mathbf{u}(t_n)$ and a time step Δt_n, we compute $\mathbf{u}_{n+1} \approx \mathbf{u}(t_n + \Delta t_n)$ by solving the implicit system

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tfrac{1}{2}\Delta t_n(\dot{\mathbf{u}}_{n+1} + \dot{\mathbf{u}}_n) = \mathbf{u}_n + M^{-1}\left(\mathbf{f} - \tfrac{1}{2}A(\mathbf{u}_{n+1} + \mathbf{u}_n)\right). \tag{1.6}$$

We advocate TR because it is the most accurate A-stable method commensurate with a second order spatial discretization, and also because it is nondissipative (some consideration is given to other linear multistep methods in section 6). This is important when solving advection-dominated problems. Another positive feature is that the local truncation error is easily estimated by repeating the time step using an explicit second order Adams–Bashforth method (AB2):

$$\mathbf{u}^*_{n+1} = \mathbf{u}_n + \Delta t_n \dot{\mathbf{u}}_n + \frac{1}{2}\Delta t_n^2\left(\frac{\dot{\mathbf{u}}_n - \dot{\mathbf{u}}_{n-1}}{\Delta t_{n-1}}\right). \tag{1.7}$$

A more subtle issue is that implemention of this linear multistep pair within a self-adaptive algorithm needs to be done carefully. Indeed, a naive implementation may well have a tendency to "stall" since rounding errors often accumulate and cause the time steps to asymptote and prevent them from increasing as they should. This is illustrated in Figure 1.1 (top) which shows the behavior of Δt_n versus t_n for the initial value problem (IVP) $\dot{y} = -0.01y$, $y(0) = 1$ with the error per step tolerance (to be defined later) $\varepsilon = 10^{-4}$ ($\times$) and $\varepsilon = 10^{-7}$ ($\circ$). (The first two time steps are $\Delta t_0 = \Delta t_1 = 10^{-10}$.) Instead of increasing to infinity with n, it is found that $\Delta t_n \to 135.1$ in the first case and $\Delta t_n \to 0.1351$ in the second, indicating that asymptotically,[1] $\Delta t_n \sim O(\varepsilon)$. The fact that this long-time behavior is spurious is

[1] This can be proven, but we do not include the proof here.

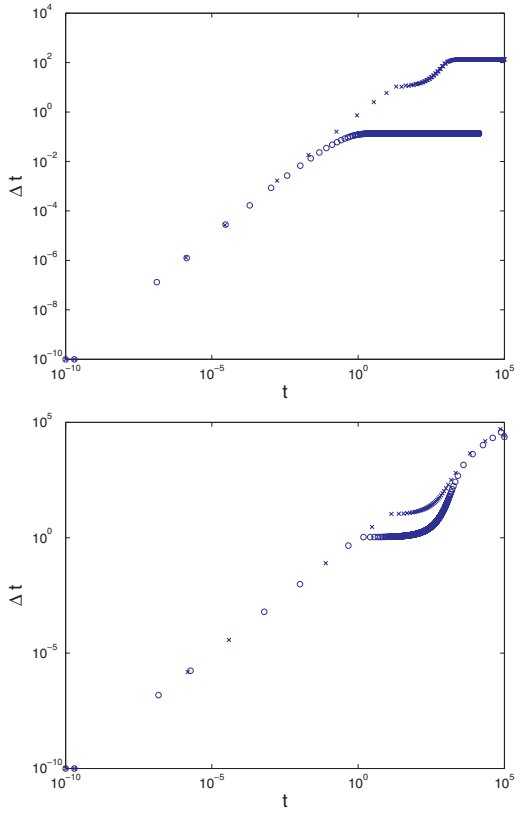

FIG. 1.1. *Top: Log-log of the time steps Δt_n vs. t for a naive TR-AB2 integration of $\dot{y} = -0.01y$ with tolerances $\varepsilon = 10^{-4}$ (×), 10^{-7} (○), and $\Delta t_0 = 10^{-10}$. Bottom: Corresponding plot using a numerically stable TR-AB2 integrator.*

confirmed in Figure 1.1 (bottom) which shows the time step histories obtained for a mathematically equivalent algorithm [14] (see also [7, p. 273]), which uses exactly the same startup time steps and tolerances. In this case, it is seen that $\Delta t_n \to \infty$ as $n \to \infty$.

Conscious of the need to minimize potential round-off instability, our implementation of the TR-AB2 pair explicitly computes the vector updates scaled by the time step to avoid underflow and inhibit subtractive cancellation. Specifically, given $\mathbf{u}_n$, $\dot{\mathbf{u}}_n$, and $\ddot{\mathbf{u}}_n$, we compute a vector $\mathbf{v}_n$ via

$$\left(M + \tfrac{1}{2}\Delta t_n A\right)\mathbf{v}_n = M\dot{\mathbf{u}}_n - A\mathbf{u}_n + \mathbf{f}, \tag{1.8}$$

and update the TR solution vector and time derivative via

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tfrac{1}{2}\Delta t_n \mathbf{v}_n; \qquad \dot{\mathbf{u}}_{n+1} = \mathbf{v}_n - \dot{\mathbf{u}}_n. \tag{1.9}$$

(The more obvious way of writing the right-hand side (RHS) of (1.8) as $-2A\mathbf{u}_n + 2\mathbf{f}$ is more prone to the ringing phenomenon discussed later in this section. The reason for this is discussed in [7, pp. 272–273].) Similarly, the scaled AB2 update $\mathbf{w}_n$ is explicitly given by

$$\mathbf{w}_n = \dot{\mathbf{u}}_n + \tfrac{1}{2}\Delta t_n \ddot{\mathbf{u}}_n \tag{1.10}$$

and generates the AB2 estimate and the second time derivative (needed for the following AB2 step) via

$$\mathbf{u}^*_{n+1} = \mathbf{u}_n + \Delta t_n \mathbf{w}_n, \qquad \ddot{\mathbf{u}}_{n+1} = \frac{\dot{\mathbf{u}}_{n+1} - \dot{\mathbf{u}}_n}{\Delta t_n}. \tag{1.11}$$

Standard manipulations, see, e.g., [7, p. 265], then lead to the truncation error estimate

$$\mathbf{u}_n - \mathbf{u}(t_n) = \frac{1}{12}\Delta t_n^3 \dddot{\mathbf{u}}(\hat{t}) \;\approx \mathbf{d}_n = \frac{\Delta t_n}{3(1+\Delta t_{n-1}/\Delta t_n)}\left(\frac{1}{2}\mathbf{v}_n - \mathbf{w}_n\right). \tag{1.12}$$

Time step selection. To control the time integration it is usual to place a user-specified *tolerance*, ε, on the norm of $\mathbf{d}_{n+1}$:

$$\|\mathbf{d}_{n+1}\| \le \varepsilon \|\mathbf{u}\|_\infty. \tag{1.13}$$

For our target problem (1.4) we use the L_2 function norm

$$\|\mathbf{d}_n\| = (\mathbf{d}_n^T M \mathbf{d}_n)^{1/2} \tag{1.14}$$

as this will ensure that Δt_n remains constant for pure advection. An appropriate choice for $\|\mathbf{u}\|_\infty$ is (a possibly user-specified estimate of) the maximum norm of the ODE solution over the prescribed time interval.[2] Assuming that our ODE system has smooth third derivatives in time (so that the TR time integration is indeed second order accurate) standard manipulation of Taylor series shows that the ratio of successive truncation errors is proportional to the cube of the ratio of successive time steps. This implies that

$$\|\mathbf{d}_n\|(\Delta t_{n+1}/\Delta t_n)^3 \lessapprox \varepsilon \|\mathbf{u}\|_\infty.$$

Thus, assuming $\|\mathbf{u}\|_\infty = 1$ and invoking equality (corresponding to taking the maximum possible time step to satisfy the accuracy tolerance at the next step) leads to the following time step selection heuristic:

$$\Delta t_{n+1} = \Delta t_n \left(\varepsilon / \|\mathbf{d}_n\|\right)^{\frac{1}{3}}. \tag{1.15}$$

To implement this methodology in a practical code there are two start-up issues that need to be addressed:

1. AB2 is not self-starting. We suggest computing $\dot{\mathbf{u}}_0 = M^{-1}(A\mathbf{u}_0 - \mathbf{f})$ and $\dot{\mathbf{u}}_1$ from (1.8) and (1.9) in order to start AB2 at the second timestep. Error control and Δt variation is then switched on at the third time step ($\Delta t_1 = \Delta t_0$).
2. Choice of initial time step Δt_0. Several strategies are available with which to start the TR method. If an estimate of the initial response time ($\tau_0 = 1/|\lambda|$, where λ is the dominant eigenvalue of the matrix $M^{-1}A$) is available, then a reasonable choice would be $\Delta t_0 = 0.01\tau_0\varepsilon^{\frac{1}{3}}$. Alternatively, one may simply select a conservatively small value for Δt_0 (say 10^{-10}). With such a choice we will have rapid growth in the time step: typically $\mathbf{d}_n = O(\texttt{eps})$ for the first few time steps, where `eps` is machine precision[3] and so $\Delta t_{n+1}/\Delta t_n =$

[2] $\|\mathbf{u}\|_\infty = \|\mathbf{u}_0\| = 1$ in all of the examples discussed in this paper.
[3] `eps` $\approx 2.22 \times 10^{-16}$ in MATLAB, which is used for all of the examples discussed in this paper.

$O((\varepsilon/\texttt{eps})^{1/3}) \approx 10^4$ when $\varepsilon = 10^{-4}$. This rapid growth implies that, for small values of n,

$$t_n = \sum_{j=0}^{n-1} \Delta t_j \approx \Delta t_{n-1},$$

and with very few such steps (typically 2–4), a time step is obtained that is commensurate with τ_0. This also explains the linear growth of Δt with t in Figure 1.1 for both implementations starting from $\Delta t_0 = 10^{-10}$. We discuss the other features of Figure 1.1 in the next section.

A general ODE code (like `ode23t`) will have many additional heuristics, bells, and whistles; see, Gresho and Sani [7, p. 275], Hairer, Norsett, and Wanner [12, p. 167], Hundsdorfer and Verwer [15, p. 197] or Shampine, Gladwell, and Thompson [24, p. 27]. Our code has just one.

1. Time step rejection. If $\|\mathbf{d}_n\| > 1.1\varepsilon$, then we consider the local error to be too large. The step is rejected, the current value of Δt_n is multiplied by $(\varepsilon/\|\mathbf{d}_n\|)^{1/3}$, and the step is repeated with this smaller value of Δt_n.

This "trip" is not really needed when solving advection-diffusion problems: in the runs reported later, rejected steps are extremely rare. The heuristic would be important, however, if the linear algebra solve in the computation of the TR update is done "inexactly" (in particular, using a preconditioned Krylov subspace solver instead of MATLAB's sparse solver). This will be the case when applying our adaptive time-stepping methodology to the Navier–Stokes equations, and we, thus, defer further discussion of rejected steps until part II which will build on the strategy outlined by Gresho and Sani [8, section 3.16.4].

Stabilization of the integrator. The solution of the IVP $\dot{y} = -\lambda y$, $y(0) = y_0$ solved using the numerically stable TR-AB2 method (as in Figure 1.1 (bottom)) can be shown to satisfy a recurrence with an explicit solution given by

$$\begin{bmatrix} y_n \\ \frac{1}{\lambda}\dot{y}_n \end{bmatrix} = \frac{y_0 + \frac{1}{2}\Delta t_0 \dot{y}_0}{1 - \frac{1}{2}\lambda\Delta t_0} \prod_{j=0}^{n-1} \frac{1 - \frac{1}{2}\lambda\Delta t_j}{1 + \frac{1}{2}\lambda\Delta t_j} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{1.16}$$

Looking at (1.16) suggests a potential problem caused by the product of rational factors. As $\lambda\Delta t_n \to \infty$ the factors for large n tend to -1 and so both y_n and $\frac{1}{\lambda}\dot{y}_n$ would behave asymptotically like $(-1)^n$—this is the familiar "ringing" phenomenon for TR. Although we have not observed this problematic behavior when solving scalar ODEs, ringing effects are often observed for very stiff PDEs (typically with very small spatial grid sizes to resolve fine detail) with relatively large tolerances on the time step or towards the end of a simulation when close to steady state. Situations such as these are discussed by Osterby [22] along with a variety of means of suppressing the oscillations. Our code implements an alternative strategy—*time step averaging*. The averaging is invoked periodically every n_* steps. For such a step, having computed the TR update $\mathbf{v}_n$ via (1.8) we set $t_{n+1} = t_n + \frac{1}{2}\Delta t_n$ and update the solution vectors via the sequence

$$\mathbf{u}_n = \tfrac{1}{2}(\mathbf{u}_n + \mathbf{u}_{n-1}); \qquad \dot{\mathbf{u}}_n = \tfrac{1}{2}(\dot{\mathbf{u}}_n + \dot{\mathbf{u}}_{n-1}); \tag{1.17}$$

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tfrac{1}{4}\Delta t_n \mathbf{v}_n; \qquad \dot{\mathbf{u}}_{n+1} = \tfrac{1}{2}\mathbf{v}_n. \tag{1.18}$$

We then compute the next time step using (1.15) and continue the integration.

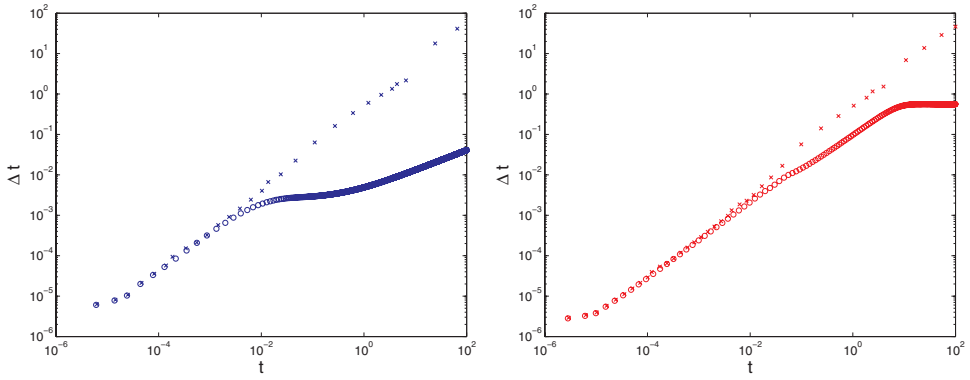

FIG. 1.2. *Left: Log-log plot of* Δt *vs.* t *for advection-diffusion of a step profile on a Shishkin grid: standard TR-AB2 integrator* ($\circ$) *and stabilized TR-AB2 integrator* ($\times$) *with tolerance* $\varepsilon = 10^{-3}$. *Right: Corresponding plot for finer tolerance* $\varepsilon = 10^{-4}$.

The averaging process annihilates any contribution of the form $(-1)^n$ to the solution and its time derivative, thus cutting short the "ringing" while maintaining second order accuracy. In our code the parameter n_* is computed automatically. We specify a target time, t_*, that is longer than the response time $\tau_0 = 1/|\lambda|$, and then set n_* to be the number of steps taken to reach this time starting from $t = 0$.[4] The benefit of this simple stabilization strategy is illustrated in Figure 1.2 which shows the behavior of stabilized and unstabilized TR-AB2 for advection-diffusion of a step profile with diffusion parameter $\nu = 10^{-3}$ and a Shishkin grid with $N = 128$ subintervals. More details of the experimental set-up are given in Example 5.2, discussed later. All time steps apart from the first two (both 10^{-10}) are shown. For both the fine (right plot) and coarse (left plot) tolerance the time steps generated by stabilized and unstabilized versions are very similar up to $t \approx 10^{-2}$. Thereafter the time steps used by the unstabilized version are smaller (in the coarse case, considerably smaller). Reducing the tolerance in the unstabilized version delays the onset of instability. The stabilized method for $\varepsilon = 10^{-4}$ (in this case with $n_* = 12$) reaches the target time in 43 rather than 249 time steps. For $\varepsilon = 10^{-3}$ the frequency of averaging is $n_* = 9$ steps and 32 steps are used (as opposed to the 4600 steps required by the unstabilized version). It may appear that the averaging process is invoked very frequently but in these experiments it is only called on three times. A more typical value of n_* with smaller tolerances is $n_* \approx 100$. The ringing effect is a consequence of the lack of L-stability in the TR scheme, see [11, p. 45], and is effectively countered by our stabilization process.

An outline of the rest of the paper is as follows. We analyze the behavior of the TR-AB2 integrator when applied to a scalar ODE model in the next section. Then, in order to demonstrate the performance of the integrator over a wide range of conditions, we discuss six example problems in detail; two pure diffusion problems, three advection-diffusion problems (all advection-dominated), and one pure advection problem. In each case we give a (mainly) qualitative analysis explaining the observed behavior of the integrator. We believe that, at the end, the reader will be strongly convinced not only that advection-diffusion problems benefit significantly by the use of an adaptive time integrator but also that studying the behavior of the time step often helps to delineate different phases of the evolution.

[4] $t_* = 10^{-4}$ in all of the examples discussed in this paper.

2. A model ODE problem. To whet the appetite, let us take a closer look at the performance of a numerically stable version of TR-AB2 (as in Figure 1.1) applied to the standard scalar ODE test equation

$$\dot{y} = -\lambda y, \qquad y(0) = 1, \tag{2.1}$$

with the solution $y(t) = e^{-\lambda t}$. The general case of $\lambda = 1/\tau + \mathrm{i}\,\omega$ is considered here. In particular, τ represents a decay time constant mimicking diffusion and the frequency ω gives a simple model for advection. From (1.16) we have $y_{n+1} = (1 - \frac{1}{2}\lambda\Delta t_n)/(1 + \frac{1}{2}\lambda\Delta t_n)y_n$ and $\dot{y}_n = -\lambda y_n$, and on substituting into the scalar analogue of (1.12) we get the explicit expression

$$d_n = -\frac{\lambda^3 \Delta t_n^3 y_n}{12(1 - \frac{1}{2}\lambda\Delta t_{n-1})(1 + \frac{1}{2}\lambda\Delta t_n)}.$$

The time step selection heuristic (1.15) then implies that

$$\Delta t_{n+1}^3 = \frac{12\varepsilon}{|\lambda^3 y_n|}\left|\left(1 - \frac{1}{2}\lambda\Delta t_{n-1})(1 + \frac{1}{2}\lambda\Delta t_n\right)\right|, \tag{2.2}$$

and we deduce that

$$\frac{\Delta t_{n+2}}{\Delta t_{n+1}} = \left|\frac{1 + \frac{1}{2}\lambda\Delta t_{n+1}}{1 - \frac{1}{2}\lambda\Delta t_{n-1}}\right|^{1/3}. \tag{2.3}$$

This is a three-step recurrence with initial conditions $\Delta t_0, \Delta t_1$ prescribed and Δt_2 given by (2.2). To make progress we distinguish the special case $\Re\lambda = 0$ from the more general case $\Re\lambda = 1/\tau > 0$.

First, the recurrence (2.3) is stable and clearly has a constant solution when $\lambda = \mathrm{i}\,\omega$ ($\omega > 0$). From (2.2) we find that

$$\Delta t_2 = \Delta t_3 = \cdots = \Delta t_n = \frac{(12\varepsilon)^{1/3}}{\omega} + O(\varepsilon/\omega). \tag{2.4}$$

Since $|y_n| = 1$, there is no amplitude error—just phase error—and the global error $|y(t_n) - y_n|$ ranges between 0 and 2, that is from perfectly in-phase to completely out-of-phase. This periodic behavior can be further analyzed by setting $y_n = \mathrm{e}^{\mathrm{i}\omega_{\Delta t} t_n}$, where $\omega_{\Delta t} = \omega - \frac{1}{12}\omega^3\Delta t^2 + O(\Delta t^4)$ is the numerical frequency, so that

$$\begin{aligned} |y(t_n) - y_n| &= \left|\mathrm{e}^{-\mathrm{i}\omega t_n} - \mathrm{e}^{-\mathrm{i}\omega_{\Delta t} t_n}\right| = 2|\sin\tfrac{1}{2}(\omega - \omega_{\Delta t})t_n| \\ &= 2|\sin\tfrac{1}{24}\omega^3\Delta t^2 t_n| + O(\Delta t^4). \end{aligned} \tag{2.5}$$

Thus, provided the time interval $[0, t^*]$ is such that $\omega^3 t^* \Delta t^2 \ll 1$, the global error is second order and grows linearly in time (this is typical behavior; see [3, ch. 9], for instance). This is also illustrated quite clearly in Figure 2.1 (top) where log-log plots of the global error are shown for tolerances $\varepsilon = 10^{-4}, 10^{-7}$ (the errors for $t = t_1, t_2$ are not shown since they are much too small). Using (2.5) leads to the simple estimate, T, of the period of the "beats" in the global error

$$T = 24\pi/\omega^3\Delta t^2.$$

For $\omega = 1$ and $\varepsilon = 10^{-4}$, we get $\Delta t \approx (12\varepsilon)^{1/3} = 0.1063\ldots$ giving $T = 6676.9$ vis-à-vis the numerical result 6675.6 in Figure 2.1 (bottom).

Second, if $\Delta t_{n+1} \geq \Delta t_{n-1}$ and $\Re\lambda > 0$, the RHS of (2.3) is > 1 and so $\Delta t_{n+2} > \Delta t_{n+1}$. Thus, by induction, the sequence $\{\Delta t_n\}$ grows monotonically. We will identify three distinct phases of time step growth in the case $\Re\lambda > 0$.

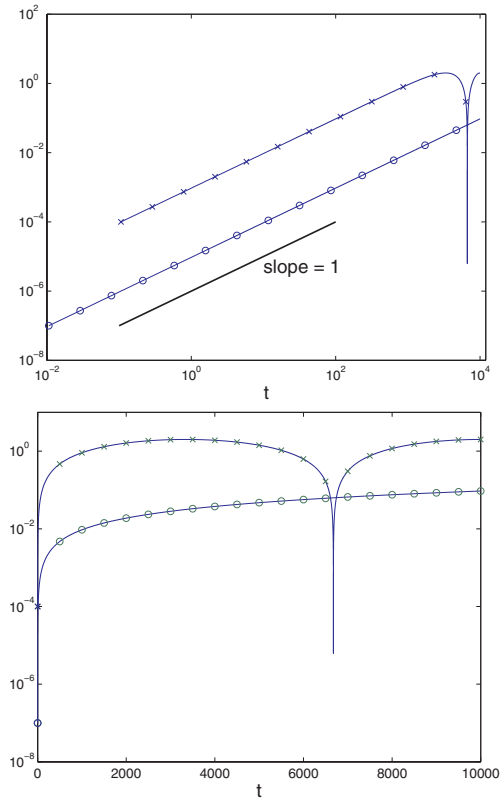

FIG. 2.1. *The global error* $|y(t_n) - y_n|$ *vs.* t *for TR-AB2 integration of* $\dot{y} = -\mathrm{i}y$*, with tolerances* $\varepsilon = 10^{-4}$ *(×) and* $\varepsilon = 10^{-7}$ *(○). Top: Log-log, Bottom: Log-linear.*

A start-up phase. As discussed earlier, see Figure 1.1, the time step rapidly increases from any (conservatively small) initial value to a value that is appropriate for the physical response time $\tau_0 = 1/|\lambda|$ and the selected tolerance. The observed behavior of Δt vs. t (growing and then flattening) can be predicted analytically, at least in the case of real λ [14].

A transient phase as the solution relaxes to its rest state. Insight into the dynamical behavior in this phase can be obtained through a modified equation approach (see [29], [10], [9]). We reparameterize time by a pseudo-time variable s discretized with constant step-size $\Delta s = (12\varepsilon)^{1/3}$. Since $\Delta t_n/\Delta s$, is an approximation to $\mathrm{d}t/\mathrm{d}s$, then (2.2) is a consistent finite difference approximation of the ODE

$$\frac{\mathrm{d}t}{\mathrm{d}s} = |\lambda|^{-1}|y|^{-1/3}. \tag{2.6}$$

Using the chain rule $\frac{\mathrm{d}y}{\mathrm{d}s} = -\lambda\frac{\mathrm{d}t}{\mathrm{d}s}y$, gives

$$\frac{\mathrm{d}y}{\mathrm{d}s} = -\lambda|\lambda|^{-1}|y|^{-1/3}y. \tag{2.7}$$

The numerical solution y_n of our time integrator and the time levels t_n are approximated by $y(s_n)$ and $t(s_n)$, the solutions of the coupled system (2.6) and (2.7). Mul-

TABLE 2.1
Actual behavior of TR-AB2 for $\dot{y} = -0.01y$.

ε	10^{-4}	10^{-7}	10^{-10}	
n (n^*)	34 (28)	292 (282)	2836 (2823)	$\sim \varepsilon^{-1/3}$
$\|y(t_n) - y_n\|_\infty$	4.71×10^{-4}	5.35×10^{-6}	5.42×10^{-8}	$\sim \varepsilon^{2/3}$

tiplying (2.7) by $\bar{y}$ and integrating gives

$$|y(s)| = \left(1 - \frac{s}{3}\left(\frac{\Re\lambda}{|\lambda|}\right)\right)^3, \tag{2.8}$$

so that, with respect to this new parameterization, the approach to the stationary point $y = 0$ is cubic rather than exponential. Solving (2.8), the steady state will be reached in the finite "pseudo-time" $s^* = 3|\lambda|/\Re\lambda$, and since $s_n = n\Delta s$ this gives a total of $n^* = (1/(12\varepsilon)^{1/3})(3|\lambda|/\Re\lambda)$ time steps. Note that n^* is independent of λ when λ is real. In practice, if we compare n^* with n (the actual number of time steps taken by our integrator to satisfy $|y_n| < \varepsilon$), then there is almost perfect agreement. A typical set of results is given in Table 2.1.

Equations (2.6) and (2.8) imply (unsurprisingly) that $\lambda t = -\log|y|$. The assumption that $|\lambda|\Delta t \ll 1$ can then be used to simplify (2.2) leading to the estimate

$$\Delta t_n \approx \frac{(12\varepsilon)^{1/3} \mathrm{e}^{\Re\lambda t_n/3}}{|\lambda|}, \tag{2.9}$$

which predicts that the time step will grow exponentially. These predicted time steps are shown by dotted curves in the top of Figure 2.2 and again the agreement with computed time steps is excellent—the predicted and actual behavior is indistinguishable to graphical accuracy for $t < 250$. We also indicate by the three vertical solid lines the times at which the numerical solution passes through $|y| = \varepsilon$. The modified equation (2.7) cannot be expected to be valid in this neighborhood (or for longer times). (2.9) coincides with the constant time step in (2.4) when $\Re\lambda = 0$.

Computed global errors are shown in the bottom of Figure 2.2 for the three values of the tolerance ε, and it is seen that the global error is reduced by a factor of 100 when ε is reduced by a factor 1000 in keeping with a global error of $O(\varepsilon^{2/3})$. The tails in Figure 2.2 oscillate when the solution reaches the level of the tolerance and thus begin when $t = O(\log(1/\varepsilon))$.

Long term behavior. Our computional experiments have been carried out over unreasonably long time intervals in order to display this behavior clearly. As was mentioned earlier, (2.3) shows that the time steps grow strictly monotonically when $\Re\lambda > 0$. The rate of increase is largest when $\Delta t_{n-1} \approx 2/\Re\lambda$.[5] From (2.9), this occurs when

$$|y_n| \approx \frac{3\varepsilon}{2}\left(\frac{\Re\lambda}{|\lambda|}\right)^3,$$

that is when $|y_n| = O(\varepsilon)$, so that the transient behavior of the true solution is well over. After this point, the time steps continue to increase by the ratio on the RHS of (2.3). This ratio tends to unity (and becomes independent of λ) as $\Delta t_n \to \infty$, thus the rate of increase in the time step is progessively reduced—as can be observed in both Figure 1.1 (bottom) and Figure 2.2 (top). Notice that, as well as being independent of λ, the long term behavior of the time step is clearly independent of ε.

[5]It is theoretically possible for the denominator on the RHS of (2.3) to vanish, in which case $y_n = 0$ and the calculation stops. We shall ignore this unlikely possibility.

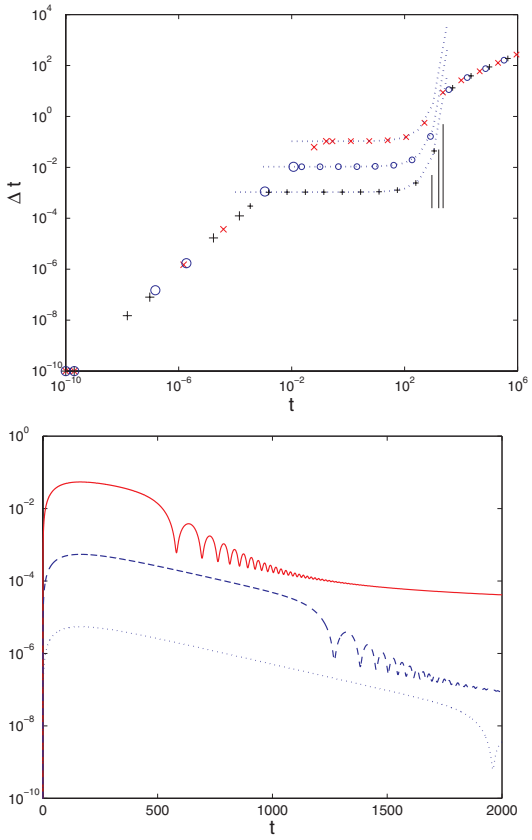

FIG. 2.2. *Top: Log-log of the time steps Δt_n vs. t_n for TR-AB2 integration of $\dot{y} = -(0.01+\mathrm{i})y$, with $\varepsilon = 10^{-4}$ (×), 10^{-7} (○), and 10^{-10} (+). Bottom: Log-linear plots of the global error against t.*

3. The heat equation. We now consider the heat equation, that is the case of $a = 0$ in (1.1). Our objective is to relate the temporal variation of the time step to the smoothness of the initial data. To start with, we assume homogeneous Dirichlet BCs in (1.2) and (1.3) and suppose that $u_0(x)$ has a Fourier sine series

$$u_0(x) = \sum_{j=1}^{\infty} a_j \sin j\pi x \tag{3.1}$$

that satisfies $\sum_1^\infty a_j^2 < \infty$ so that it is convergent for $u_0 \in L_2(0,1)$. When the Fourier coefficients decay more quickly than this, specifically, $\sum_{j=1}^\infty j^{2\theta} a_j^2 < \infty$, for some $\theta > 0$, then we say that $u_0 \in H_0^\theta(0,1)$, and we define a norm on this space by

$$|u_0|_\theta^2 = \frac{1}{2}\sum_{j=1}^{\infty} j^{2\theta} a_j^2. \tag{3.2}$$

See Babuška and Strouboulis [2, pp. 113–129] for a more complete discussion of these spaces and the concept of square integrable fractional derivatives. The solution of the heat equation with this initial data can be expressed as

$$u(x,t) = \sum_{j=1}^{\infty} a_j \mathrm{e}^{-\nu j^2 \pi^2 t} \sin j\pi x. \tag{3.3}$$

For later reference we bound the norm of u_{ttt}. If we assume that $u_0 \in H^\theta(0,1)$ $(0 \le \theta < 6)$, then, by Parseval's relation,

$$\begin{aligned}
\|u_{ttt}\|_0^2 &= \frac{1}{2}\nu^6\pi^{12}\sum_{j=1}^{\infty} a_j^2 j^{12}\mathrm{e}^{-2\nu j^2\pi^2 t} \\
&\le \frac{1}{2}\nu^6\pi^{12}(\max_j \mathrm{e}^{-2\gamma\nu j^2\pi^2 t})\sum_{j=1}^{\infty} a_j^2 j^{2\theta} j^{12-2\theta}\mathrm{e}^{-2(1-\gamma)\nu j^2\pi^2 t} \\
&\le \frac{1}{2}\nu^6\pi^{12}\mathrm{e}^{-2\gamma\nu\pi^2 t}\big(\max_j j^{12-2\theta}\mathrm{e}^{-2(1-\gamma)\nu j^2\pi^2 t}\big)\sum_{j=1}^{\infty} a_j^2 j^{2\theta} \\
&= \nu^6\pi^{12}\mathrm{e}^{-2\gamma\nu\pi^2 t}\big(\max_j j^{12-2\theta}\mathrm{e}^{-2(1-\gamma)\nu j^2\pi^2 t}\big)|u_0|_\theta^2 \qquad (3.4)
\end{aligned}$$

for any $\gamma \in [0,1)$.[6] Now, $\max_j j^{2\alpha}\mathrm{e}^{-\beta j^2} = (\alpha/\beta)^\alpha\,\mathrm{e}^{-\alpha}$, from elementary calculus, leading to the estimate

$$\|u_{ttt}\|_0 \le \pi^6\nu^3\left(\frac{6-\theta}{2(1-\gamma)\nu\mathrm{e}\pi^2 t}\right)^{3-\theta/2}\mathrm{e}^{-\gamma\nu\pi^2 t}|u|_\theta. \qquad (3.5)$$

The decay in time in (3.5) is associated with the concept of "parabolic smoothing" (see, for example, [4], [18], [19]): $\|u_{ttt}\|_0 \le Ct^{-(3-\theta/2)}\mathrm{e}^{-\gamma\nu\pi^2 t}$ for $t > 0$, C being a constant depending on ν, γ, and θ. For small times the algebraic decay dominates whereas decay is governed by the exponential factor (with γ arbitrarily close to, but less than one) in the long run. In the case of very smooth initial data, $u_0 \in H^\theta(0,1)$ for $\theta \ge 6$, the maximum in (3.4) occurs at $j = 1$ and we obtain

$$\|u_{ttt}\|_0 \le \pi^6\nu^3\mathrm{e}^{-\nu\pi^2 t}|u|_\theta, \qquad (3.6)$$

an exponential decay with a rate that is dictated by the smallest eigenvalue $(\nu\pi^2)$ of the diffusion operator $-\nu u_{xx}$; this is independent of θ.

We now turn to the analysis of the semidiscrete approximation of the heat equation with homogeneous Dirichlet BCs (as in (1.4))

$$M\dot{\mathbf{u}} = -K\mathbf{u}, \qquad (3.7)$$

where M and K are symmetric positive-definite tridiagonal matrices. We suppose that the generalized eigenvalue problem

$$M\mathbf{v} = \lambda K\mathbf{v} \qquad (3.8)$$

has eigenvalues ordered so that $\lambda_1 < \lambda_2 < \cdots < \lambda_{N-1}$ and associated eigenvectors normalized so that $\mathbf{v}_j^T M\mathbf{v}_j = 1$. The solution of (3.7) can then be written as

$$\mathbf{u}(t) = \sum_{j=1}^{N-1} c_j\mathrm{e}^{-\lambda_j t}\mathbf{v}_j, \qquad (3.9)$$

where the coefficients $\{c_j\}$ are determined from the initial data by $c_j = \mathbf{v}_j^T M\mathbf{u}(0)$.

[6] The introduction of γ is purely a technical device—its aim is to give the exponential decay at large times by choosing γ close to, but less than one. It could be avoided by redoing the analysis with $\gamma = 0$ with the proviso that the results are only useful at short times. The exponential phase can then be handled separately, taking $u \sim a_1\mathrm{e}^{-\pi^2 t}\sin\pi x$ and the spatial and temporal errors from the leading terms in (3.20) and (3.22), respectively.

When the TR-AB2 integator is applied to the system (3.7) there are three distinct time scales in the evolution; two of these are directly related to the eigenvalues λ_j and the third to parabolic smoothing. These are discussed in turn below.

Fast transient. At early times the variation in the solution is dominated by the fastest transient in (3.9): $\mathbf{u}(t) \approx c_N \mathrm{e}^{-\lambda_{N-1} t}\mathbf{v}_{N-1} +$ slower varying terms. Thus, as in the scalar case (see (2.9)), we have that

$$\Delta t_{n+1} \approx \frac{(12\varepsilon)^{1/3}}{|c_{N-1}|^{1/3}\lambda_{N-1}} \mathrm{e}^{\lambda_{N-1} t_n/3}. \tag{3.10}$$

Since the high frequency modes in the numerical solution (3.9) bear no relation to the corresponding modes in the PDE solution, this phase is spurious; the numerical solution cannot begin to approximate the true solution unless all of the coefficients of high frequency modes are sufficiently small (which occurs if the initial data has a high degree of smoothness) or have sufficiently decayed.

Smoothing phase. Given the definition of the temporal truncation error, (1.12), we suppose the existence of a constant C such that the bound $\|\mathbf{d}_n\| \le \frac{1}{12} C \Delta t_n^3 \|u_{ttt}\|_0$ holds. Combining the time step selection heuristic (1.15) with the bounds (3.5) (with $\gamma = 0$) and (3.6) then leads to the estimate

$$\Delta t_{n+1} \ge \frac{1}{C} \times \frac{(12\varepsilon)^{1/3}}{|u|_\theta^{1/3}\nu\pi^2} \times \begin{cases} \left(\frac{2\nu \mathrm{e} \pi^2 t_n}{6-\theta}\right)^{1-\theta/6} & \text{if } \theta \le 6 \\ \mathrm{e}^{\nu\pi^2 t_n/3} & \text{if } \theta > 6. \end{cases} \tag{3.11}$$

The lower bound suggests that the time step grows sublinearly for $\theta < 6$ and grows exponentially for $\theta > 6$.

Relaxation to steady state. As $t \to \infty$, the solutions of the heat equation and its spatial approximation are governed by the low frequency eigenvectors corresponding to the smallest eigenvalues: $\mathbf{u}(t) \approx c_1 \mathrm{e}^{-\lambda_1 t}\mathbf{v}_1$. Thus, as in the scalar case (see (2.9)), we have that

$$\Delta t_{n+1} \approx \frac{(12\varepsilon)^{1/3}}{|a_1|^{1/3}\nu\pi^2} \mathrm{e}^{\nu\pi^2 t_n/3}. \tag{3.12}$$

This asymptotic form is more precise than (3.11) for long times in the case $\theta > 6$.

As for the ODE model discussed in section 2 there could be two additional (nonphysical) phases of time step growth: the startup phase as Δt_n grows from its initial value, and the long term behavior which is independent of ε.

We now give an example that illustrates that these estimates of the time step can be realized in actual computations.

Example 3.1. Consider the system (3.7) arising from discretizing $u_t = u_{xx}$ on $0 < x < 1$ with BCs $u(0,t) = 1$, $u(1,t) = 0$, and an initial condition given by

$$u_0(x) = 1 - |1 - 2x|^\alpha, \qquad \alpha > 0. \tag{3.13}$$

Note that this is rough initial data for all $\alpha > 0$. In addition to the obvious discontinuities in derivatives at $x = 1/2$ when α is not an even integer, there are singularities since the initial condition does not satisfy the higher order compatibility conditions required by the Dirichlet boundary conditions. When u_0 is extended as an odd function to the interval $(-1, 1)$, appropriate for homogeneous Dirichlet boundary conditions, there is a discontinuity in the second derivative at the origin and so the coefficients $\{a_j\}$ cannot decay more quickly that $1/j^3$. Moreover, in view of the dis-

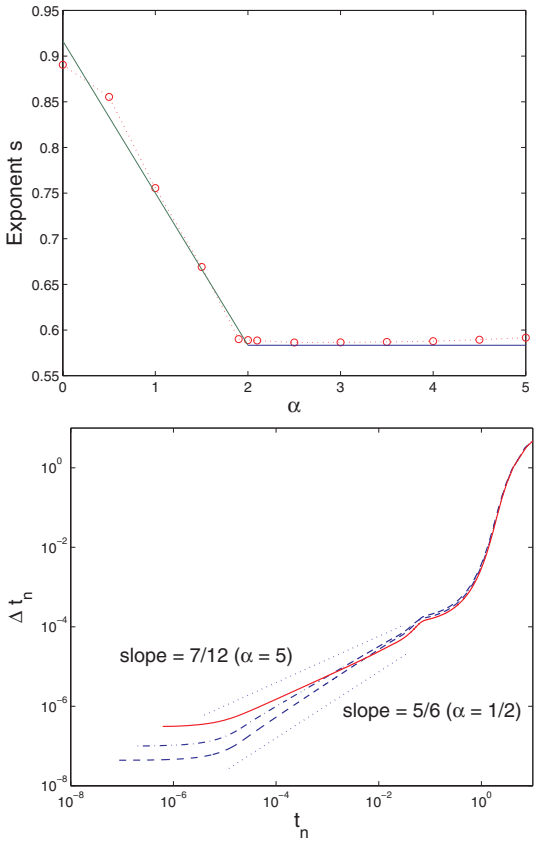

FIG. 3.1. *Top: The observed rate of growth of time steps for Example* 3.1 *(∘) and the rate predicted by* (3.14) *(solid). Bottom: Log-log plot of time steps* Δt *vs.* t *for* $\alpha = \frac{1}{2}$ *(broken line),* $\alpha = 1$ *(dot-dashed line) and* $\alpha = 5$ *(solid line). The predicted rates for* $\alpha = \frac{1}{2}, 5$ *are shown by dotted lines.*

continuities in derivatives at $x = \frac{1}{2}$ for $\alpha < 2$, $u_0 \in H^{p-\delta}(-1,1)$ for any $\delta > 0$, where $p = \max\{\alpha + \frac{1}{2}, \frac{5}{2}\}$.[7] Given this level of regularity, the estimate (3.11) predicts that during the smoothing phase the time step will be bounded below by

$$\Delta t_{n+1} \geq C t^s, \qquad s = \max\{7/12, 11/12 - \alpha/6\}, \tag{3.14}$$

where the constant C depends on α and $|u_0|_{p-\delta}$.

We take a subdivision of $N = 256$ equal elements and use a small tolerance $\varepsilon = 10^{-10}$ in order to accurately determine the observed rates of growth—obtained by computing the slopes of the linear regression lines through the values of $\log \Delta t_n$ versus $\log t_n$ for $t_n \in [5 \times 10^{-5}, 5 \times 10^{-3}]$. The resulting exponents s are shown on the top of Figure 3.1 while on the bottom we show log-log plots of Δt_n against t_n for $\alpha = \frac{1}{2}, 1, 5$. The agreement between theory and practice is excellent.[8]

[7] The presence of δ is a consequence of using the norm (3.2) on H^θ. It could be removed by using a (more appropriate) Besov space; see [2].

[8] In order to get this level of agreement, spatial resolution is not as important as having a small tolerance; increasing ε to 10^{-7} leads to quite poor agreement while decreasing the number of elements to 64 has little effect (except when $\alpha = 0$, which we interpret as the discontinuous function $u_0(x) = 1 - |1 - 2x|\text{sign}(1 - 2x)$).

Our second example introduces some other important behavioral characteristics without the added complication of advection (which will be included in the next section).

Example 3.2. Consider the system (1.4) arising from discretizing $u_t = \nu u_{xx}$ on $0 < x < 1$ with $\nu = 1$, BCs $u(0,t) = 1$, $u(1,t) = 0$, and an initial condition given by $u_0(x) = 1$, $0 \le x \le 1$.

This PDE problem is particularly challenging because of the impossibility of obtaining a solution of "arbitrary" accuracy for $x \to 1$ and $t \downarrow 0$, owing to the singularity there. At early times, $0 < t < 10^{-2}$, we compare our numerical solution with the classical error function approximation to the solution

$$u_{\mathrm{erf}}(x,t) = \mathrm{erf}\big((1-x)/\sqrt{4\nu t}\big). \tag{3.15}$$

In particular, using Maple 9.5©[9] we can show that

$$\left\| \frac{\partial^3 u_{\mathrm{erf}}}{\partial t^3} \right\|^2 = \int_0^1 \left(\frac{\partial^3}{\partial t^3} \mathrm{erf}\left(\frac{1-x}{\sqrt{4\nu t}} \right) \right)^2 \mathrm{d}x \sim \frac{945}{2048} \frac{\sqrt{2\nu/\pi}}{t^{11/2}},$$

where $\sim$ indicates that this is accurate up to exponentially small terms of the form $\exp(-1/\nu t)$. Then, assuming $\Delta t \approx \left(12\varepsilon / \| \frac{\partial^3 u_{\mathrm{erf}}}{\partial t^3} \| \right)^{1/3}$, this leads to the estimate

$$\Delta t(t) \sim (12\varepsilon)^{1/3} \left(\frac{2048}{945\sqrt{2\nu/\pi}} \right)^{1/6} t^{11/12}. \tag{3.16}$$

Note that the step function initial data has a regularity estimate $u_0 \in H_\theta(0,1)$ for $\theta < 1/2$, see [2, p. 126], so the factor $t^{11/12}$ in the time step growth in (3.16) is completely consistent with the bound in (3.11).

For $t \ge 10^{-2}$ our numerical solution will be compared with the truncated Fourier series (FS) solution:[10]

$$u_n(x,t) = 1 - x + \sum_{j=1}^{n} \frac{2}{j\pi} \exp(-j^2\pi^2\nu t) \sin j\pi x. \tag{3.17}$$

We turn next to the issue of spatial resolution. A notional definition of the thickness of the growing boundary layer at small times is, from (3.15), given by $\delta(t) := \sqrt{4\nu t}$ ($\delta(0.01) = 0.02$ at the time when (3.15) is replaced by (3.17)). We combine this with the concept of *minimum time of believability*, see Gresho and Sani [7, p. 196], which is defined by

$$\tau_{\mathrm{MTB}} := h^2/4\nu.$$

This is the time at which the boundary layer width has grown from zero to approximately h, the distance to the first node from the boundary i.e., $\delta(\tau_{\mathrm{MTB}}) = h$. Clearly, no numerical approximation can be accurate for $\delta(t) \ll \delta(\tau_{\mathrm{MTB}})$. For $\delta(t) > \delta(\tau_{\mathrm{MTB}})$ on the other hand—that is for $t > \tau_{\mathrm{MTB}}$—believability becomes at least plausible.

For a 256-element uniform mesh, the above remarks are validated by the computational results shown in Figure 3.2 where we show both the analytical solution

[9] Copyright (C) Maplesoft, a division of Waterloo Maple Inc.

[10] For $0 < t < 10^{-2}$, $\|u - u_{\mathrm{erf}}\|_\infty < 10^{-12}$, and we can ensure comparable accuracy at later times $\|u - u_n\|_\infty < 10^{-12}$ by taking $n = \lceil 5/\pi\sqrt{\nu t} \rceil$ terms of the FS.

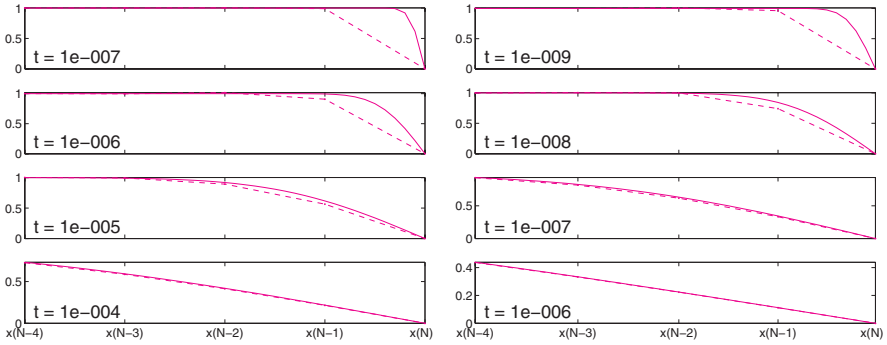

FIG. 3.2. *Numerical solutions (broken lines) and exact solution (solid lines) in the four elements nearest the boundary where the singularity is present. Left: Uniform grid. Right: Geometric grid.*

(solid line) and the FE solution (broken lines) at four values of t. In this case we have $\tau_{\text{MTB}} \approx 3.8 \times 10^{-6}$, and from the figure we see that the numerical solution does not resolve the boundary layer before $t \approx 10^{-5}$. We shall analyze the global error behavior in more detail in a moment.

Here, in order to make our numerical solution useful for much smaller times (while still retaining its usefulness for larger times, of course) we now introduce a 256-element "smart" (or at least, smarter) mesh. To do this we select—admittedly somewhat arbitrarily—a small time believability limit of $\tau_{\text{MTB}} = 10^{-8}$ (nearly 400 times smaller than with the uniform mesh) and use the definition of τ_{MTB} to generate the smallest grid size $h = h_{\min}$ on our new mesh ($h_{\min} = 2 \times 10^{-4}$, nearly 20 times smaller than h for the uniform mesh). The remaining nodal locations come from the geometric formula

$$h_j = \rho^{N-j} h_{\min}, \quad j = 1 : N,$$

with the grid ratio ρ chosen so that $\sum_{j=1}^{N} h_j = 1$. With $N = 256$ this gives $\rho \approx 1.0177$ and a largest element at $x = 0$ with $h_{\max} \approx 0.0176$—88 times larger than $h_{\min}$. The solutions at early times in the four elements closest to the singularity are also shown in Figure 3.2. The numerical solution is (almost) useful by $t = \tau_{\text{MTB}} (= 10^{-8})$ and the superiority of this grid is clear.

In Figure 3.3 we show, on a log-log scale, how the time step varies for four cases, two on the uniform mesh and two on the geometric mesh, all starting with $\Delta t_0 = 10^{-10}$. The first point we wish to emphasize is the tremendously large variation in step sizes—eleven orders of magnitude in the "best" case (geometric mesh with $\varepsilon = 10^{-7}$) using about 850 time steps (see Table 3.1). Clearly, there is no fixed-Δt integrator that could even begin to compete with this efficient use of time steps! The next thing to note is the numerical agreement with theory: a thousandfold change in ε gives a tenfold change in step size. In each case the time steps adjust from the initial value $\Delta t_0 = 10^{-10}$ to a more (ε-dependent) appropriate value in no more that four steps.

We have included in Figure 3.3 three vertical broken lines that delineate the main phases of the evolution. The first occurs at $t = \tau_{\text{MTB}}$: at earlier times the time step behaves as given by the fast-transient expression (3.10) (shown by a dotted curve for $\varepsilon = 10^{-4}$—the agreement is much better when $\varepsilon = 10^{-7}$). The interval $\tau_{\text{MTB}} < t < \tau_1$ ($\tau_{\text{MTB}} \approx 4 \times 10^{-6}$ for the uniform grid and 10^{-8} for the geometric grid while $\tau_1 = O(1/\lambda_1) \approx 1/\nu\pi^2$) covers the smoothing phase when the solution is

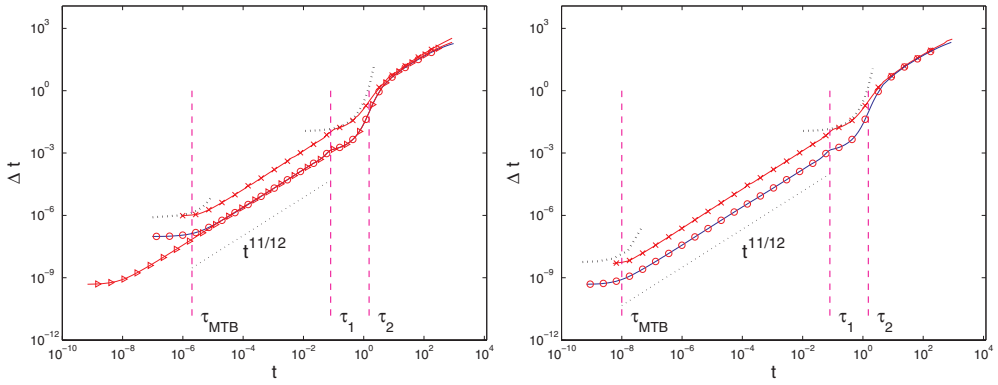

FIG. 3.3. *Log-log plot of time steps Δt vs. t for Example* 3.2. *Left: Uniform grid. Right: Geometric grid. The grids have $N = 256$ elements, $\varepsilon = 10^{-4}$ ($\times$), and $\varepsilon = 10^{-7}$ ($\circ$). Also, shown on the left are the time steps for the geometric grid ($\triangleright$) for comparison.*

TABLE 3.1
Total number of steps used by uniform and geometric grids in Example 3.2 *for* $0 < t \leq 10$.

Grid		$\varepsilon = 10^{-4}$	$\varepsilon = 10^{-7}$	$\varepsilon = 10^{-10}$
Uniform	$N = 128$	100	702	6647
	$N = 256$	103	743	7098
Geometric	$N = 128$	113	842	8073
	$N = 256$	114	853	8168

erf–like and the time step is given by the lower bound in (3.11) with $\theta = 1/2$ (or (3.16)). Comparing with the dotted line which has gradient 11/12, it is seen that the exponent given for the time step is sharp. The time steps for the geometric and uniform grids for $\varepsilon = 10^{-7}$ (Figure 3.3, left, $\triangleright$ and $\circ$, respectively) are virtually identical for $t > 10^{-5}$. The constant τ_2 in Figure 3.3 is the time when the solution is within $O(\varepsilon)$ of the steady state. During the third interval $\tau_1 < t < \tau_2$, the time step is governed by (3.12) (shown again by a dotted curve for $\varepsilon = 10^{-4}$). In the final interval $t > \tau_2$, the time step is governed by the internal dynamics of the TR-AB2 integrator and behaves as in the scalar case. Most notably Δt is independent of ε and the spatial grid; see the discussion at the end of section 2. It is seen in Table 3.1 that increasing the number of elements for a given tolerance has a negligible effect on the number of time steps required to integrate up to $t = 10$. This suggests that the solution is well resolved spatially. It is also seen that the use of a geometric grid costs an additional 10%–20% time steps but in return the solution is accurately obtained for much smaller times.

We now consider the behavior of the global error. We do this back-to-front: we will first discuss the computational results before deriving analytic bounds for the spatial and temporal errors consistent with the observed behavior. Figure 3.4 shows the maximum error as a function of time for both meshes (and two values of ε) and for 127- and 255-term Fourier series. The results show the effectiveness of the combination of a smart time integrator with a smart mesh—it even beats the Fourier series up until $t \approx 10^{-5}$. (Of course, a *really* smart mesh would both move and remove nodes as time progressed, ending with just 3 nodes.) For each of these meshes the behavior of the error goes through several phases. For the uniform mesh the log-log plot of the error shown on the top of Figure 3.4 initially behaves like t^{-1}. Also, when N is

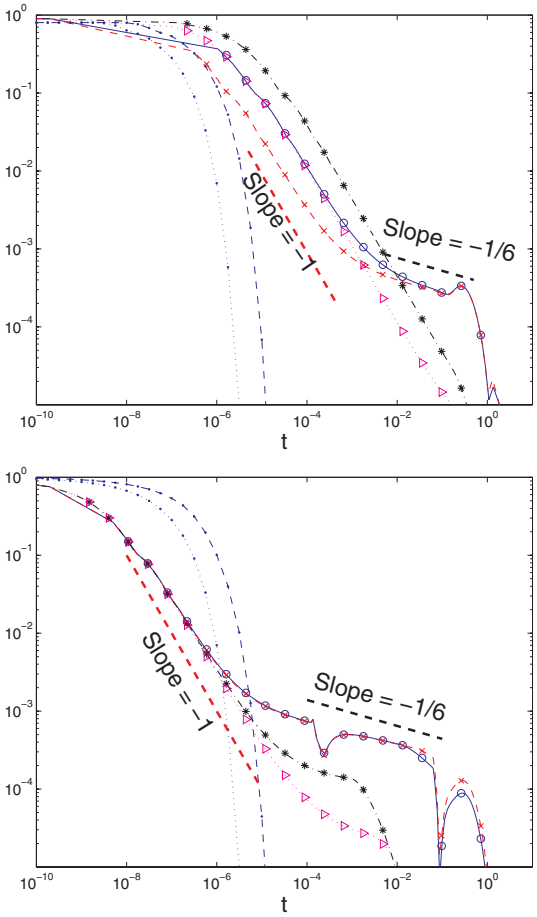

FIG. 3.4. *The maximum error as a function of time for the experiments of Example* 3.2. *Top: Uniform grid. Bottom: Geometric grid. Key:* $\varepsilon = 10^{-4}$ *(*$N = 256$ *(*$\circ$*) and* $N = 512$ *(*$\times$*)),* $\varepsilon = 10^{-7}$ *(*$N = 128$ *(*$*$*) and* $N = 256$ *(*$\triangleright$*)). Also shown are the errors for the truncated Fourier series solutions* (3.17) *with* $n = 127$ *terms (dot-dash) and* $n = 255$ *terms (*$\cdot$ *and dotted curve).*

increased from 128 through 256 to 512, the error is reduced by a factor of four each time, consistent with a second order spatial approximation. This behavior turns out also to be completely consistent with our estimate of the spatial error contribution; see (3.21) below which shows that this phase of evolution is dominated by spatial error. The flattening of the error curves for $\varepsilon = 10^{-4}$ corresponds to the fact that temporal error becomes increasingly dominant. Our estimate of the temporal error (3.22) suggests that the error should decrease like $t^{-1/6}$ in this phase, consistent with what is observed. Note that for $\varepsilon = 10^{-7}$ spatial error is always dominant. All of the geometric meshes used have $h_{\min} = 2 \times 10^{-4}$. These also display an early phase where spatial error dominates (though the gradient appears to be marginally greater than -1). As temporal error dominates it appears to behave as $t^{-1/6}$. In all of the experiments there is a final stage where the global error decays exponentially as the solution relaxes to steady state.

To theoretically analyze the global error we first need to recall the local truncation

error. Specifically, we take the standard uniform grid estimate

$$T_j^{n+1/2} = -\frac{1}{12}\left[\nu h^2 \frac{\partial^4 u}{\partial x^4} + \Delta t^2 \frac{\partial^3 u}{\partial t^3}\right]_j^{n+1/2} + \dots ; \tag{3.18}$$

see, for example, Morton and Mayers [21, p. 30]. We then take our estimate of the global error to be the function $e(x,t)$ that solves the "correction equation":

$$e_t = \nu e_{xx} - \frac{1}{12}\nu h^2 \frac{\partial^4 u}{\partial x^4} - \frac{1}{12}\Delta t^2 \frac{\partial^3 u}{\partial t^3}. \tag{3.19}$$

Note that here Δt is a function of t.

We shall determine bounds on the maximum norms of the spatial and temporal errors separately, and, to keep things simple (see (3.11)), we assume that $\theta \le 6$. (Later we are only really interested in the particular case of $\theta = 1/2$.) With u given by the Fourier expansion (3.3) the spatial error component, $e^{(S)}$, will be governed by

$$e_t^{(S)} = \nu e_{xx}^{(S)} - \frac{1}{12}\nu\pi^4 h^2 \sum_{j=1}^{\infty} j^4 a_j \mathrm{e}^{-\nu\pi^2 j^2 t} \sin j\pi x$$

from which we deduce, since $e^{(S)}(x,0) = 0$,

$$e^{(S)}(x,t) = -\frac{1}{12}\nu\pi^4 h^2 t \sum_{j=1}^{\infty} j^4 a_j \mathrm{e}^{-\nu\pi^2 j^2 t} \sin j\pi x, \tag{3.20}$$

and hence

$$\|e^{(S)}(\cdot,t)\|_\infty \le C\nu h^2 t \sum_{j=1}^{\infty} j^4 |a_j| \mathrm{e}^{-\nu\pi^2 j^2 t},$$

where C denotes a generic constant independent of $x, t, h, \Delta t$. Assuming that $u_0 \in H^\theta(0,1)$, we obtain (with any $\delta > 0$ and with $0 \le \gamma < 1$)

$$\begin{aligned}\|e^{(S)}(\cdot,t)\|_\infty &\le Ch^2(\nu t)\big(\max_j \mathrm{e}^{-\gamma\nu\pi^2 j t}\big) \sum_{j=1}^{\infty} (j^\theta |a_j|) j^{-1/2-\delta} \left(j^{9/2-\theta+\delta} \mathrm{e}^{-(1-\gamma)\nu\pi^2 j^2 t}\right) \\ &\le Ch^2(\nu t)\mathrm{e}^{-\gamma\nu\pi^2 t} \max_{j\ge 1}\left(j^{9/2-\theta+\delta} \mathrm{e}^{-(1-\gamma)\nu\pi^2 j^2 t}\right) \sum_{j=1}^{\infty} (j^\theta |a_j|) j^{-1/2-\delta}.\end{aligned}$$

Then, since $\sum_{j=1}^{\infty} j^{-1-2\delta} < \infty$, using the Cauchy–Schwarz inequality gives

$$\|e^{(S)}(\cdot,t)\|_\infty \le C\nu h^2 |u_0|_\theta \frac{\mathrm{e}^{-\gamma\nu\pi^2 t}}{(\nu t)^{5/4-\theta/2-\delta/2}}. \tag{3.21}$$

For practical purposes, we can set $\gamma = 1$ and $\delta = 0$. Thus, in the case of interest $\theta = 1/2$, our bound suggests that the spatial error initially behaves like t^{-1} and then ultimately decays exponentially.

The temporal error component $e^{(T)} := e - e^{(S)}$ is governed by

$$e_t^{(T)} = \nu e_{xx}^{(T)} - \frac{1}{12}\Delta t^2 \nu^3 \pi^6 \sum_{j=1}^{\infty} j^6 a_j \mathrm{e}^{-\nu\pi^2 j^2 t} \sin j\pi x$$

from which we deduce, since $e^{(T)}(x,0)=0$, that

$$e^{(T)}(x,t) = -\frac{1}{12}\Delta t^2 \nu^3 \pi^6 t \sum_{j=1}^{\infty} j^6 a_j \mathrm{e}^{-\nu\pi^2 j^2 t} \sin j\pi x.$$

We now assume that $\Delta t(t)$ is given by the lower bound in (3.11) for $\theta < 6$, that is

$$e^{(T)}(x,t) = C\,(\varepsilon/|u_0|_\theta)^{2/3}\,(\nu t)^{3-\theta/3} \sum_{j=1}^{\infty} j^6 a_j \mathrm{e}^{-\nu\pi^2 j^2 t} \sin j\pi x.$$

Then, the same argument that was used to estimate the spatial error gives the estimate

$$\|e^{(T)}(\cdot,t)\|_\infty \le C\varepsilon^{2/3}|u_0|_\theta^{1/3} \frac{\mathrm{e}^{-\gamma\nu\pi^2 t}}{(\nu t)^{1/4-\theta/6-\delta/2}}. \tag{3.22}$$

The algebraic powers of t in (3.22) and (3.21) are equal when $\theta = 3$. For $\theta < 3$ and for early times the bounds suggest that the spatial error is dominant. In the case of interest $\theta = 1/2$, the bound (3.22) is consistent with the behavior seen in Figure 3.4, that is, for dominant temporal error the global error behaves like $t^{-1/6}$.

In some recent related works, Verfürth [27] has presented *a posteriori* energy error estimates of theta time stepping for the fully discretized heat equation, and Akrivis, Makridakis, and Nochetto [1] give refined *a posteriori* error estimates for TR semidiscretization of selfadjoint parabolic equations (which does not include convection-diffusion or a discussion of spatial discretization). The nature of corner singularities for the heat equation and their effects on numerical simulations have been studied by Flyer and Fornberg [6].

4. Pure advection of a smooth wave form. We now take a rather large step—from a "rough" diffusion problem to a "smooth" advection problem—in part to see what different aspects of the TR-AB2 might surface. The smoothness of the initial data assures good spatial resolution with relatively few elements on a uniform grid. The solution is governed by the ODE system (1.4) with $\nu = 0$, and if periodic boundary conditions had been considered, $A = C$ would then be a circulant skew-symmetric matrix. In such a case, as in (3.9), the solution would be written in the form

$$\mathbf{u}(t) = \sum_{j=1}^{N} c_j \mathrm{e}^{-\lambda_j t} \mathbf{v}_j, \tag{4.1}$$

except that now λ_j are imaginary (cf. (2.4)). This would, in turn imply that $\|\mathbf{u}\| \equiv (\mathbf{u}^T M \mathbf{u})^{1/2}$ was conserved in time, as would the time derivatives of the solution: $\|\dot{\mathbf{u}}\|$, $\|\ddot{\mathbf{u}}\|$, and $\|\dddot{\mathbf{u}}\|$, which would further imply a constant sequence of time steps

$$\Delta t_2 = \Delta t_3 = \cdots = \Delta t_n \approx \frac{(12\varepsilon)^{1/3}}{\|\dddot{\mathbf{u}}\|^{1/3}}. \tag{4.2}$$

Returning to the general (nonperiodic) case, the global error e is defined by the following analogue of (3.19),

$$e_t + a e_x = T(x,t) = -\frac{1}{12}\Delta t^2 \frac{\partial^3 u}{\partial t^3} - \frac{1}{180} a h^4 \frac{\partial^5 u}{\partial x^5}, \tag{4.3}$$

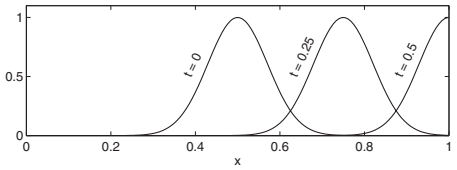

FIG. 4.1. *Pure advection of Gaussian given by* (4.5) *at* $t = 0, 0.25, 0.5$.

where $T(x,t)$ is the truncation error term. Since $u \equiv u(x-at)$ and Δt is constant, it follows that $T \equiv T(x-at)$ and that (4.3) has the solution

$$e(x,t) = t\,T(x-at), \tag{4.4}$$

which implies that the error grows linearly with time along characteristics $x - at =$ constant. This behavior of the error will also be seen in the following example.

Example 4.1. Consider the system (1.4) arising from discretizing $u_t + u_x = 0$ on $0 < x \le 1$, with BC $u(0,t) = 0$ for $t \ge 0$, and the initial condition given by a discrete version of the Gaussian profile

$$u_0(x) = \exp\{-(x-x_0)^2/2\sigma^2\}, \tag{4.5}$$

which is centered at $x_0 = 1/2$ and which has $\sigma = 1/\sqrt{200} \approx 0.071$.

The semidiscrete equation at $x = 1$ is given by

$$\tfrac{1}{6}h(2\dot{U}_N + \dot{U}_{N-1}) + \tfrac{1}{2}(U_N - U_{N-1}) = 0, \tag{4.6}$$

which means that K is no longer skew-symmetric. Figure 4.1 shows the numerical solution at several times for $N = 128$. Clearly the well-resolved Gaussian is easily tracked through these grids. The TR-AB2 time step histories are shown on the top in Figure 4.2 for two values of N, namely 128 and 256, and two values of the tolerance $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-7}$. These time steps can be compared with the "theoretical values" obtained by replacing $\ddot{\mathbf{u}}$ in (4.2) by $\partial^3 u_\infty/\partial t^3$, where

$$u_\infty(x,t) = \exp(-(x-x_0-t)^2/2\sigma^2) \tag{4.7}$$

is the travelling wave solution that would arise on an infinite span—this prediction is graphically indistinguishable from the computed values shown up to a time of $t \approx 5/6$, when the Gaussian has effectively left the domain. Indeed, for $0 < t \lessapprox 0.4$ the time step is constant (to machine precision for $\varepsilon \lessapprox 10^{-5}$) and may be estimated (in a similar way to (3.16)) from the full span Gaussian to be

$$\Delta t_\infty = (12\varepsilon)^{1/3}\left(8\sigma^5/(15\sqrt{\pi})\right)^{1/6}, \tag{4.8}$$

which gives $\Delta t_\infty \approx 9.6 \times 10^{-3}$ for $\varepsilon = 10^{-4}$ and is a factor of 10 smaller for $\varepsilon = 10^{-7}$, in agreement with Figure 4.2. Also, up to the time $t \approx 5/6$ the time steps differ imperceptibly for the two values of N which indicates that the spatial error is negligible.

The global error $e = u - U$ for $0 < t < 0.25$ is shown in Figure 4.3 with $\varepsilon = 10^{-7}$ and $N = 128$ and grows according to the predicted form (4.4). The analogous figures for the other parameter values scale as $\varepsilon^{2/3}$ since the global error is dominated by temporal error in this time interval.

We return now to the discussion of the time step histories in Figure 4.2. The exact solution (that is u, not u_∞) is virtually zero for $t > 1$ which would lead one to

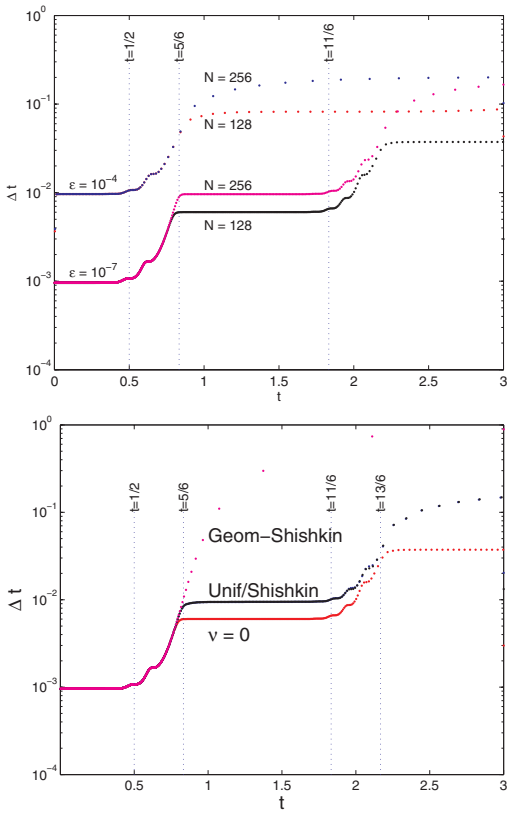

FIG. 4.2. *Time step history for propagation of a Gaussian. Top: Example* 4.1, *pure advection on uniform grids with* $\varepsilon = 10^{-7}$ *and* $\varepsilon = 10^{-4}$. *Bottom: Example* 5.1, *advection-diffusion with* $\nu = 2 \times 10^{-5}$ *and* $\varepsilon = 10^{-7}$.

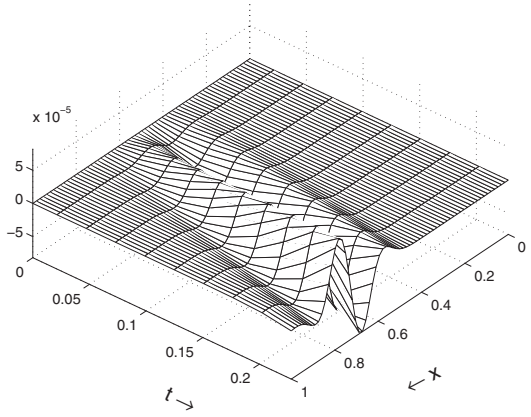

FIG. 4.3. *Global error for pure advection of a Gaussian for* $0 \le t \le 0.25$, $\varepsilon = 10^{-7}$ *and* $N = 128$.

suppose that the time steps should increase to infinity. However, the computed time steps are approximately constant for $5/6 < t < 11/6$. The problem, not apparent on the scale of Figure 4.1, is that the numerical solution has some difficulty in leaving the computational domain cleanly. This is manifested in a very small trail of "wiggles" (with wave lengths $\approx 2h$) that move upstream at the group velocity (-3) of a $2h$ wave (see Gresho and Sani [7, section 2.6]). The wiggles are a consequence of the outflow equation and their properties can be deduced from the semidiscrete equations. We analyze them in some detail since they have a significant bearing on the examples in the next section.

To isolate the reflections we need to compare the FE solution U with the FE solution U_∞ on the domain $\{(x,t) : -\infty < x < \infty, t > 0\}$ with initial data extended to infinity by zero. However, U_∞ is an $O(h^4)$ approximation to u_∞, the solution of the advection equation on the same domain. Since the reflections are of $O(h^2)$ it suffices to look at the truncation error associated with the outflow approximation (4.6) which gives

$$\begin{aligned} T_N &= \tfrac{h}{6}(2\dot{u}_N + \dot{u}_{N-1}) + \tfrac{1}{2}(u_N - u_{N-1}) \\ &= \tfrac{1}{2}h[u_t + u_x]_N - \tfrac{1}{12}h^2[2u_{xt} + 3u_{xx}]_N + \cdots . \end{aligned} \tag{4.9}$$

The reflection induced by T is given by the solution, say ρ, of the semidiscrete equations (1.5) (with $\nu = 0$) with $\rho_j(0) = 0$ for $0 < j \le N$, $\rho_0(t) = 0$ and the outflow equation

$$\tfrac{h}{6}(2\dot{\rho}_N + \dot{\rho}_{N-1}) + \tfrac{1}{2}(\rho_N - \rho_{N-1}) = -\tfrac{1}{12}h^2 u_0''(1-t), \tag{4.10}$$

where the RHS has been obtained by substituting the exact solution $u_\infty = u_0(x-t)$ into the above expression for T. Since the initial data u_0 is only defined on the interval $(0,1)$, the RHS of this equation is nonzero only for $0 < t < 1$.

We now assume that the values of ρ at even and odd numbered grid points are approximations to separate smooth functions $p(x,t)$ and $q(x,t)$. The internal semidiscrete equations will then be consistent of order $O(h^2)$ with the hyperbolic system

$$\left.\begin{aligned} \tfrac{1}{3}(2p_t + q_t) + q_x = 0 \\ \tfrac{1}{3}(p_t + 2q_t) + p_x = 0 \end{aligned}\right\} \tag{4.11}$$

while (4.10) leads to

$$\tfrac{1}{6}h(2p_t + q_t) + \tfrac{1}{2}(p - q) = -\tfrac{1}{12}h^2 u_0''(1-t).$$

To leading order $p = q = O(h^2)$ so the first term on the left-hand side (LHS) is $O(h^3)$ which may be neglected. This leads to the outflow BC

$$p - q = -\tfrac{1}{6}h^2 u_0''(1-t) \tag{4.12}$$

at $x = 1$ and the RHS is again zero for $t > 1$. Adding and subtracting the component equations in (4.11) reveals that $p + q$ and $p - q$ are constant along the characteristic lines $x - t =$ constant and $x + 3t =$ constant, respectively (the slopes of these lines correspond to the group speeds of 1 for long wavelengths and -3 for $2h$-wavelengths).

We focus first on the boundary values of the solution. Since $p = q = 0$ at $t = 0$ and $p + q = 0$ on the incoming characteristic at $x = 1$, the BC (4.12) gives

$$p(1,t) = -q(1,t) = -\tfrac{1}{12}h^2 u_0''(1-t), \quad 0 < t < 1. \tag{4.13}$$

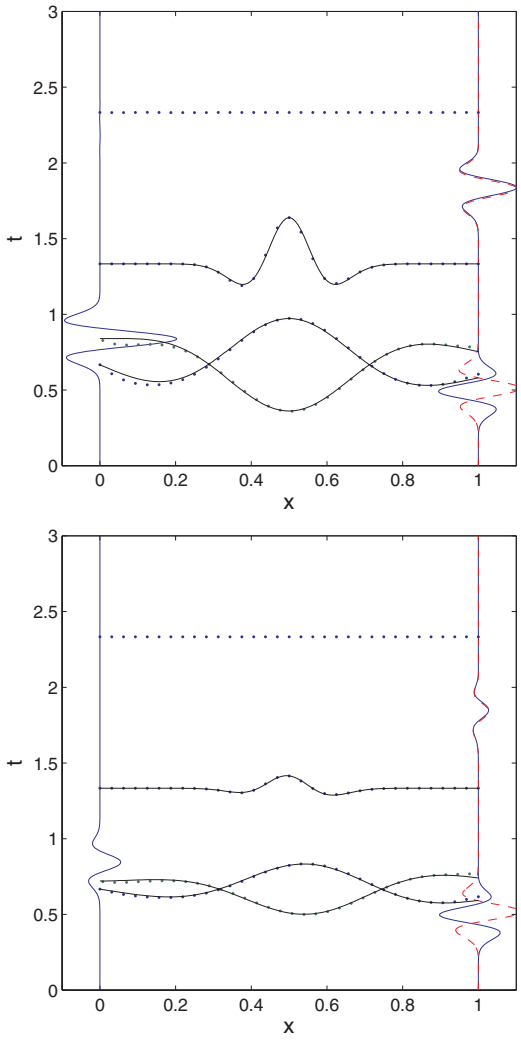

FIG. 4.4. *Plots of p and q (scaled by a factor of* 100*) showing reflections caused by the outflow equation (*$N = 128$*) for pure advection (top) and with added diffusion (*$\nu = 2 \times 10^{-5}$*) (bottom). The horizontal curves at* $t = 2/3, 4/3, 7/3$ *show the numerical error (dots) and predicted behavior (solid lines).*

This data is swept into the domain along the left-going characteristic along which $p - q$ is equal to its value at $x = 1$. Since $p = 0$ at $x = 0$, we then find that

$$q(0,t) = \tfrac{1}{6}h^2 u_0''(\tfrac{4}{3} - t), \quad \tfrac{1}{3} < t < \tfrac{4}{3}$$

so that the amplitude of q at the left boundary is twice its value at the right boundary. This boundary data is then carried back into the domain along right-going characteristics along which $p + q$ is constant and the process repeats for $4/3 < t < 7/3$. (Our asymptotics break down on this "rebound" because the forcing data on the RHS of (4.12) is zero for $t > 1$ and higher order terms have to be included.) The computed values of $U_1(t)$ (corresponding to $q(0,t)$) and $U_{N-1}(t)$, $U_N(t)$ (corresponding to $q(1,t)$ and $p(1,t)$), for $N = 128$ and $\varepsilon = 10^{-7}$ are shown as vertical curves in Figure 4.4 (top). These curves have been amplified by a factor of 100 and the chosen tolerance

is sufficiently small so that spatial errors dominate. The predicted curves are graphically indistinguishable from the computed values. The horizontal curves in Figure 4.4 (top) show the computed solutions (dots) at times $t = \frac{2}{3}, \frac{4}{3}, \frac{7}{3}$ (only alternate even and odd points are shown) and the predicted reflections (solid lines).

This explains what is going on in Figure 4.2. The first (left-going) reflection does not affect the growth of Δt since $\|\ddot{\mathbf{u}}\|$ is still dominated by the tail of the exiting Gaussian. The vertical dotted lines in Figure 4.2 are the times the right-going characteristic passing through the peak of the initial profile and its consequent reflections hit the endpoints. Following the reflection from $x = 0$, the time step is constant $\Delta t = O((\varepsilon N^2)^{1/3})$ for $5/6 \lessapprox t \lessapprox 11/6$. At $t = 1$ with $\varepsilon = 10^{-7}$ the ratio of time steps with $N = 128$ and 256 is 1.598 compared with the predicted ratio of $4^{1/3} = 1.587$. Subsequent reflections have amplitude $O(h^4)$ and so $\Delta t = O((\varepsilon N^4)^{1/3})$.

5. Advection-diffusion. We are now ready to consider the advection-diffusion equation (1.1) with $u(0,t)$ specified as inflow BC and, as in the previous section, initial data that are smooth on the open interval $(0,1)$. This precludes, for instance, the study of travelling fronts if they have regions of large gradient. We shall begin with the easier case of a "natural" outflow BC $u_x(1,t) = 0$ before proceeding to a "hard" BC $u(1,t) = 0$. The former is perhaps of greatest interest because of the difficulty in achieving the latter in physical situations, however, the latter has become a benchmark for computations because of the numerical difficulties it presents.

5.1. Natural outflow boundary conditions. With smooth, compactly supported initial data u_0, a weak boundary layer develops gradually at the outflow $x = 1$. Expanding in the parameter ν/a we find that the solution of (1.1) under these conditions can be approximated by

$$\bar{u}(x,t) = u_\infty(x,t) - \frac{\nu}{a}\mathrm{e}^{-a(1-x)/\nu}\frac{\partial u_\infty}{\partial x}(x,t), \tag{5.1}$$

where the truncation error term is $O((\nu/a)^2)$, and where u_∞ denotes the corresponding infinite span solution with the same initial data extended to be zero outside $(0,1)$, as in section 4.

If $\nu/a \ll 1$, then the numerical solution is prone to spurious reflections from the right-hand boundary as in the pure advection case discussed above. With the diffusion term present, the analogues of the hyperbolic system (4.11) and the boundary equation (4.12) are

$$\left.\begin{array}{rcl} \frac{1}{3}(2p_t + q_t) + aq_x &=& -\frac{2\nu}{h^2}(p-q) \\ \frac{1}{3}(p_t + 2q_t) + ap_x &=& \frac{2\nu}{h^2}(p-q), \end{array}\right\} \tag{5.2}$$

and

$$p - q = T_N = \tfrac{2}{a}[\nu u_x - \tfrac{1}{12}ah^2 u_{xx}]_N + \ldots,$$

respectively, where T_N is the truncation error at the outflow (cf. (4.9)). Following the approach in section 4 we substitute the truncated estimate of the solution $\bar{u}(x,t)$ into the above expression for T_N and then solve the system (5.2) with initial data $p = q = 0$ and boundary datum $p = 0$ on $x = 1$. This leads to the conclusion that $p+q$ is constant along characteristics $x - at =$ constant, but that there is exponential decay,

$$p - q = T_N\left(\frac{x-1}{3a}\right)\mathrm{e}^{-\frac{4\nu}{ah^2}(1-x)}, \tag{5.3}$$

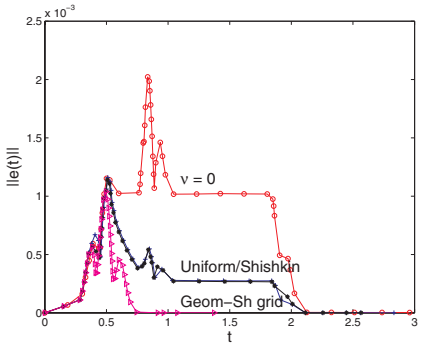

FIG. 5.1. *Global error vs. t for advection-diffusion of a Gaussian.*

along left-going characteristics $x + 3at =$ constant.

Example 5.1. Consider the system (1.4) arising from discretizing $u_t + au_x = \nu u_{xx}$ on $0 < x \leq 1$, with BCs $u(0,t) = 0$ and $u_x(1,t) = 0$ for $t \geq 0$, and the initial condition given by the Gaussian profile (4.5) as in Example 4.1.

Predicted reflections (solid lines) and the numerical results for a uniform grid with $N = 128$ (dots showing alternate even and odd grid values) for the case $a = 1$, $\nu = 2 \times 10^{-5}$ are also shown in Figure 4.4. Comparing these results with those with $\nu = 0$ the level of exponential decay is gentle but noticeable. In order to inhibit these reflections we clearly require much better resolution of the outflow boundary layer.

Perhaps the simplest way of achieving this increased resolution is to use a so-called Shishkin grid [20, 23]. In such a grid $N/2$ elements are equally spaced in each of the subintervals $[0, 1-\beta]$ and $[1-\beta, 1]$, where the "boundary layer thickness" is given by,

$$\beta = \min\left(\frac{1}{2}, \frac{2\nu}{a}\ln N\right). \tag{5.4}$$

For advection-dominated problems, $\beta = (2\nu/a)\ln N$. Using such a Shishkin grid with $N = 256$—so that the coarse grid is essentially the same as for the previous experiment—does not dampen the wiggles for reasons given below. In computations the results are graphically indistinguishable from those in Figure 4.4. This can also be seen by comparing the second and third rows of Table 5.1 where we give the magnitudes of the reflected waves at $x = 1/2$, for the times $t = 2/3, 4/3, 8/3$.

Figure 5.1 shows the behavior of the maximum norm of the global error[11] for the cases reported in Table 5.1. In all cases the error is approximately linear in time up to $t \approx 0.2$ (as in (4.4) for pure advection), and grows thereafter up to $t \approx 0.5$ (when the peak of the Gaussian meets the right boundary). The error for pure advection is then nearly constant up to $t \approx 1.7$ except for two peaks near $t = 5/6$ when the wave reflects from the left boundary, and it finally drops to zero at $t \approx 2.1$. In contrast, for both the uniform grid ($N = 128$) and the Shishkin grid ($N = 256$) when diffusion is present, the error decays on $0.5 < t < 5/6$ in accordance with (5.3) before mirroring the behavior seen in the case of pure advection.

Further insight is provided by Figure 5.2 where the computed solution and the global errors are plotted for $t = 0.4$. It can be seen that the uniform grid ($N = 128$,

[11] When computing the global error we approximate $u(x,t)$ by the $O((\nu/a)^2)$ approximation (5.1) which makes use of the infinite span solution $u_\infty(x,t) = \frac{\exp(-(x-x_0-at)^2/(2\sigma^2+4\nu t))}{\sqrt{1+2\nu t/\sigma^2}}$.

TABLE 5.1
Amplitude of the reflected waves at $x = 1/2$ for Gaussian initial data.

	ν	N	$t = 2/3$	$t = 4/3$	$t = 8/3$
Pure advection	0	128	1.023×10^{-3}	-1.019×10^{-3}	-6.76×10^{-8}
Uniform grid	2×10^{-5}	128	0.524×10^{-3}	-0.269×10^{-3}	-4.83×10^{-8}
Shishkin grid	2×10^{-5}	256	0.526×10^{-3}	-0.270×10^{-3}	-6.57×10^{-8}
Geometric grid	2×10^{-5}	256	1.174×10^{-5}	0.75×10^{-5}	1.32×10^{-10}

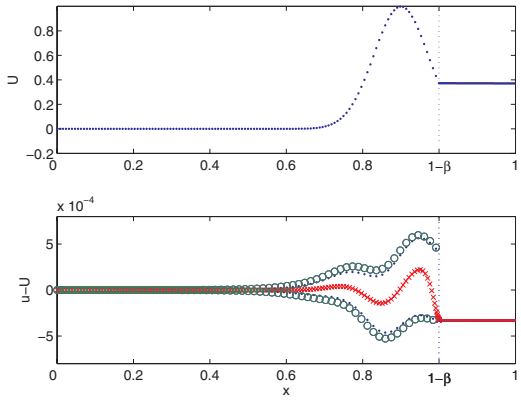

FIG. 5.2. *Top: The numerical solution for advection-diffusion of the Gaussian at $t = 0.4$ with $\nu = 2 \times 10^{-5}$ for a Shishkin grid ($N = 256$, dots). Bottom: Global error with a uniform grid ($N = 128$, circles), Shishkin grid ($N = 256$, dots), and Geometric grid ($N = 256$, crosses). The interval $(1 - \beta, 1)$ has been expanded in order to show the layer solution.*

circles) and the Shishkin grid ($N = 256$, dots) behave in essentially the same manner for $0 < x < 1 - \beta$ showing wiggles of roughly equal amplitude. These results suggest that the reason that the Shishkin grid generates reflected waves is the sharp transition between the coarse and fine grid sizes—such behavior has been known for some time for pure advection; see, for example [17].

Within the numerical layer $(1-\beta, 1)$ the magnitude of $\dot{U} = O(a)$ while the spatial derivatives are well approximated and are of $O(a/\nu)$. It can then be shown that the analogue of (5.1) is

$$U_j(t) = U_{N/2}(t) + (1 - \beta - x_j)/a\dot{U}_{N/2} + \nu e^{-a(1-x_j)/\nu}/a^2\dot{U}_{N/2} + O((\nu/a)^2) \tag{5.5}$$

for $j = N/2 : N$. Using this expansion it can be shown that the semidiscrete equation holding at the interface corresponds to the weak implementation of the BC $au_x = -u_t$ at $x = 1 - \beta$.

To reduce the reflections from the interface we adopt a "smarter" grid that is defined by

$$h_j = \tfrac{1}{2}(H + h) + \tfrac{1}{2}(H - h)\tanh\alpha(\tfrac{1}{2}(N + 1) - j), \qquad j = 1 : N \tag{5.6}$$

with $\alpha = \log(5/4)$, and satisfies $h < h_j < H$, where H and h are, respectively, the coarse and fine grid sizes used in the Shishkin grid. Since h_j approaches its extreme values geometrically (as $j \to \pm\infty$), we refer to this as a *geometric-Shishkin* grid. The largest ratio of consecutive grid sizes for the Shishkin grid is H/h while it is $e^{\alpha} = 5/4$ for this alternative grid. The grid size sequence $\{h_j\}$ is shown in Figure 5.3

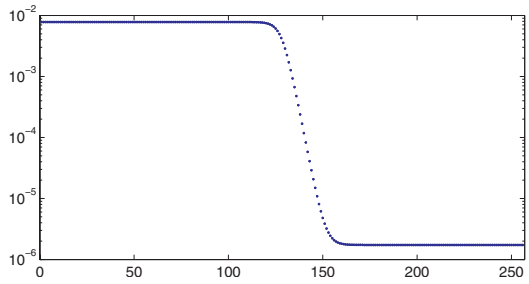

FIG. 5.3. *Grid size h_j for geometric-Shishkin grid* (5.6) *for $\nu = 2 \times 10^{-5}$ and $N = 256$.*

for $\nu = 2 \times 10^{-5}$ and $N = 256$, when we have $\beta \approx 2.2 \times 10^{-4}$ and 111 grid points are located in the boundary layer region $(1-\beta, 1)$. The global error on this geometric grid at $t = 0.4$ is plotted with crosses in Figure 5.2. It is seen that the amplitude of the error for $x < 1-\beta$ is reduced by a factor of about two, but no wiggles are discernable. The time history of the global error in Figure 5.1 confirms this fact—the plateau is no longer apparent. This better resolution of the physics is also reflected in the time step behavior using the TR-AB2 integrator. This is illustrated in Figure 4.2. Up until the Gaussian meets the right boundary the time step sequence is independent of the grid used. Thereafter Δt is dependent on the amplitude of reflected waves and so it is appreciably larger using the geometric grid compared to the Shishkin and uniform grids. Although the Shishkin grid solution can be shown to converge uniformly in ν as $N \to \infty$ [23], our geometric grid is superior (at least for these parameter values).

We have dwelt for some time on the oscillations caused at or near the outflow with the natural BC employed with the finite element method (FEM). These oscillations are insignificant when compared with the common finite difference treatment of Neumann boundary conditions using the so-called image point method. The superiority of the FEM for natural boundary conditions for advection-diffusion (or Navier–Stokes, for that matter) is clearly shown by Gresho and Sani [7, p. 209].

5.2. Dirichlet outflow boundary conditions. Our final scenario involves solving the full advection-diffusion equation (1.1) with smooth initial data $u_0(x)$ and boundary conditions $u(0,t) = u_0(0)$ and $u(1,t) = 0$ so that there is a step discontinuity at $x = 1, t = 0$. This scenario is much more challenging than that of the previous section because here it is the solution u, rather than the flux νu_x, that changes by $O(1)$ over the width $O(\nu/a)$ of the layer region. Verhulst [28] provides a brief introduction to singular perturbation techniques for situations such as this.

We consider two example problems. In the first of these we revisit Example 3.2 with the addition of advection. The effects of advection are confined to the outflow region and this allows a study of the transition from diffusion-dominated to advection-dominated flow—what we shall refer to as the advection-diffusion time scale (τ_{AD})—in its simplest setting. Our second example then looks at a more complex transition.

Example 5.2. Consider the system (1.4) arising from discretizing $u_t + a u_x = \nu u_{xx}$ on $0 < x \leq 1$, with BCs $u(0,t) = 1$ and $u(1,t) = 0$, and the step initial condition $u_0(x) = 1$, $0 \leq x < 1$.

We take $N = 256$ and concentrate on the classical Shishkin grid introduced in the previous section. Such grids are, of course, designed for exponential layers and are not ideally suited to the narrower erf-like layer that will arise at early times. Results obtained for the geometric-Shishkin grid will be discussed subsequently.

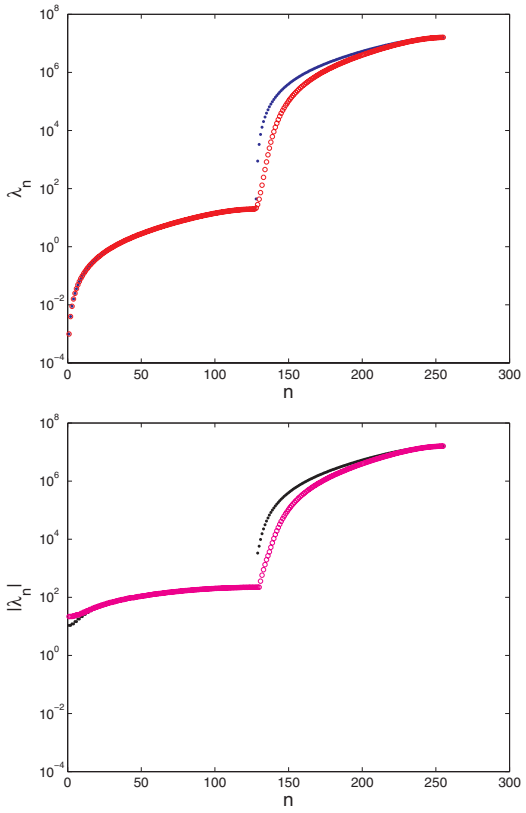

FIG. 5.4. *Eigenvalues for pure diffusion problem (top) and advection-diffusion (bottom) both with* $\nu = 10^{-4}$ *and* $N = 256$. *Symbols: Dots (•) correspond to the Shishkin grid and circles (○) correspond to the geometric-Shishkin grid.*

For comparison purposes we first discuss the results for the heat equation ($a = 0$) when using a Shishkin grid with $N = 256$ that is defined with $\beta = 2\nu \ln N$. The same problem on a uniform grid was discussed in Example 3.2. The solution of the PDE for early times is given by (3.15) and relaxes to the steady state $u(x,t) = 1 - x$ in a time $t = O(1/\nu)$.

The numerical solution, in contrast, evolves in two separate phases. In the first phase the erf-layer develops entirely within the layer $(1-\beta, 1)$ and then, in the second phase, the effect of the singularity spreads to the coarse grid on $(0, 1-\beta)$.

An understanding of this evolution requires knowledge of the associated generalized eigenvalue problem (3.8). The eigenvalues are shown in the top of Figure 5.4. They form two distinct sets, $S_c = \{\lambda_j : j = 1 : N/2 - 1\}$, corresponding to the coarse grid and $S_f = \{\lambda_j : j = N/2 + 1 : N - 1\}$ corresponding to the fine grid, with $\lambda_{N/2}$ a "bridge" between the two sets (its eigenvector has a quite different structure to those corresponding to S_c and S_f). The eigenvalues in these sets are closely approximated by the eigenvalues of the FE discretization of the differential operator $-\nu u''$ on the domains $(0, 1-\beta)$ and $(1-\beta, 1)$, respectively, each with Dirichlet boundary conditions (formulae for the discretized operator may be deduced from the results given in Gresho and Sani [7, p. 190] or [5]). Thus, we have the estimates (confirmed by

numerical experimentation)

$$\lambda_1 \approx \nu\pi^2, \lambda_{N/2-1} \approx 3\nu N^2, \lambda_{N/2} \approx \frac{\sqrt{3}N}{2\ln N}, \lambda_{N/2+1} \approx \frac{\pi^2}{4\nu(\ln N)^2}, \lambda_{N-1} \approx \frac{3N^2}{4\nu(\ln N)^2}$$

that are valid when $\nu N \ll 1$; they show the considerably different time scales on which these modes operate. (The estimate for $\lambda_{N/2}$ is conjectured from the results of numerical experimentation.)

Both phases of the numerical evolution have similarities with those in Example 3.2 with suitably modified time scales. Each has its own minimum time of believability:

$$\tau_{\mathrm{MTB}(h)} = h^2/4\nu, \qquad \tau_{\mathrm{MTB}(H)} = H^2/4\nu$$

associated with the fine (h) and the coarse (H) grid, respectively. We note that these times are also related to the largest eigenvalues in the sets S_f and S_c: $\tau_{\mathrm{MTB}(h)} = O(1/\lambda_{N-1})$ and $\tau_{\mathrm{MTB}(H)} = O(1/\lambda_{N/2-1})$.

After the first four time steps, Δt settles into the fast transient mode appropriate for the fine grid, i.e., $\Delta t \approx C\exp(\lambda_{N-1}t/3)$ for $t < \tau_{\mathrm{MTB}(h)} \approx 1.9 \times 10^{-7}$ (see Figure 5.5, top, $\circ$). The solution then enters the parabolic smoothing phase during which Δt increases as $t^{11/12}$ until $t \approx \tau_1$. At this stage all but the last of the "fast modes" from S_f have decayed and, for $\tau_1 < t < \tau_2$, $\Delta t \approx C\exp(\lambda_{N/2+1}t/3)$ based on the smallest eigenvalue from S_f. This suggests that $\tau_1 = O(1/\lambda_{N/2+1})$, where $\lambda_{N/2+1} \approx \nu\pi^2/\beta^2$.

An estimate of τ_2 may be made as the time at which the width of the singular layer (which grows proportional to $\sqrt{\nu t}$) equals the width of the fine grid region: $\sqrt{\nu t} \approx \beta$, i.e., $\tau_2 \approx \beta^2/\nu \approx 0.012$ so $\tau_2 \approx 10\tau_1$. The effect of the singularity subsequently spreads into the coarse grid and a second "fast" transient stage begins where $\Delta t \approx C\exp(\lambda_{N/2-1}t/3)$ for $\tau_2 < t < \tau_{\mathrm{MTB}(H)} \approx 0.15$. The solution then enters another smoothing stage ($\Delta t \approx Ct^{11/12}$) followed by relaxation to steady state (the time steps corresponding to these later times are not shown as the steady state is not achieved until $t = O(1/\nu)$). Overall, the time step follows the pattern of Figure 3.3 twice in succession.

In Figure 5.5 we show a linear-log plot of $\|U - u_{\mathrm{erf}}\|_\infty$ with time, where u_{erf} is the solution given by (3.15). (The maximum norm of the difference was computed by interpolating u_{erf} onto a finer grid—it is not based on just the nodal errors.) For the heat equation ($\circ$) it is seen that the norm decreases and is small ($\approx 10^{-2}$) at $t = 10^{-6}$—this suggests that the estimate $\tau_{\mathrm{MTB}(h)} \approx 1.9 \times 10^{-7}$ is a little too small. The norm continues to decrease until $t = \tau_1 \approx 10^{-3}$ when the erf layer begins to penetrate the coarse grid. During this process, which takes place in the interval $\tau_1 < t < \tau_{\mathrm{MTB}(H)}$, it grows appreciably before decaying again. The norm shows a similar rise and fall for the geometric-Shishkin grid but it is two orders of magnitude smaller. The final increase in the norm for $t > 10^3$ is due to the fact that u_{erf} does not satisfy the boundary condition $u(0,t) = 1$.

Figure 5.5 also shows a linear-log plot of $\|U - U_{\mathrm{SS}}\|_\infty$ with time, where $U_{\mathrm{SS}} = 1 - x$ denotes the steady state solution of the ODE system when $a = 0$. The solution is within 10^{-5} of the steady state by $t = 10^4 = O(1/\nu)$. Also shown by a broken line is the maximum difference $U - U_{\mathrm{SS}}$ computed over only those nodal points lying in the fine Shishkin grid. This shows that the numerical solution becomes close to steady state relatively quickly in the fine grid ($t \approx 1$).

The main difference when using a geometric-Shishkin grid—see Figure 5.5 (top, $*$ and broken line)—is that Δt behaves as $Ct^{11/12}$ for $t > \tau_{\mathrm{MTB}(h)} \approx 1.9 \times 10^{-7}$ and

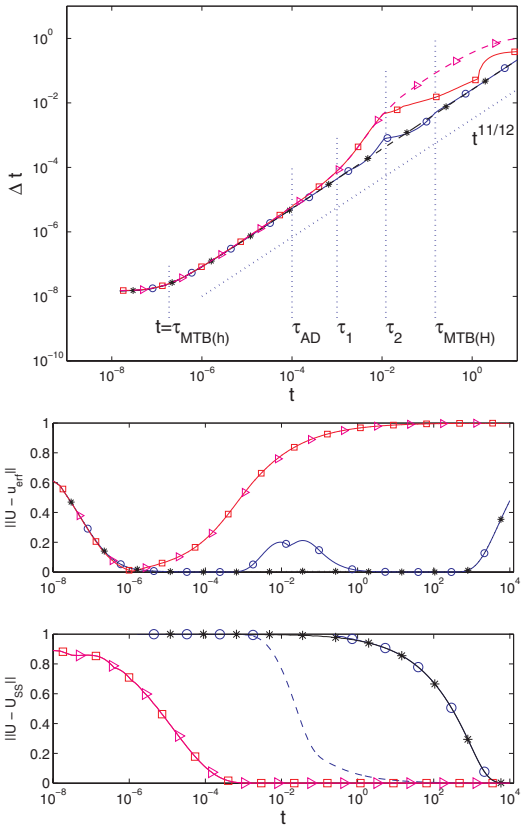

FIG. 5.5. *Top: Time step histories for Example* 5.2 *with* $\nu = 10^{-4}$*, tolerance* $\varepsilon = 10^{-7}$*, and initial time step* $\Delta t_0 = 10^{-10}$*. Bottom: Top:* $\|U - u_{\mathrm{erf}}\|_\infty$ *vs. t. Bottom:* $\|U - U_{\mathrm{SS}}\|_\infty$ *vs. t; the broken line corresponds to* $\|U - U_{\mathrm{SS}}\|_\infty$ *measured over grid points in* $(1-\beta, 1)$ *for the Shishkin grid. Key: Heat equation* ($a = 0$)*, Shishkin grid* ($\circ$)*, heat equation* ($a = 0$)*, geometric-Shishkin grid* ($*$)*, AD equation* ($a = 1$)*, Shishkin grid* ($\square$)*, AD equation* ($a = 1$)*, and geometric-Shishkin grid* ($\triangleright$)*.*

does not deviate from this, as the Shishkin grid does, over the interval $(\tau_1, \tau_{\mathrm{MTB}(H)})$. We attribute this to the fact that the eigenvalues λ_n (see Figure 5.4, top, shown by $\circ$) are more evenly distributed when n is close to $N/2$.

We now discuss the results of the experiments with advection included ($a = 1$). At early times the continuum problem is diffusion-dominated; an erf layer forms as described above for the heat equation and the solution is given by $u \approx u_{\mathrm{erf}}(x, t)$ (cf. (3.15); see, for instance, Flyer and Fornberg [6]). For long times, the solution tends to the steady state

$$u_{\mathrm{SS}}(x) = \frac{1 - \mathrm{e}^{-a(1-x)/\nu}}{1 - \mathrm{e}^{-a/\nu}} \tag{5.7}$$

having an exponential boundary layer with thickness $O(\nu/a)$. The advection-diffusion equation satisfies a maximum principle so this transition is monotonic.

The eigenvalues of the discrete advection-diffusion operator on a Shishkin grid again form two distinct sets S_c and S_f; those in S_c are complex and so it is their moduli that are shown on the bottom of Figure 5.4 ($\circ$)—the vertical scale used is the same as that in the top figure. The larger eigenvalues in S_f are closely approximated by those

of the discrete eigenvalue problem on $(1-\beta, 1)$ with homogeneous Dirichlet boundary conditions and are comparable in magnitude with those for the heat equation. This suggests that the time scales associated with exponential phases based on the fine grid are also comparable. In particular, the largest eigenvalues are roughly the same so that $\tau_{\mathrm{MTB}(h)}$ is the same for both problems. The moduli of the eigenvalues in S_c are considerably larger than the corresponding eigenvalues of the heat equation though these have little bearing on the solution due to ill conditioning; see Trefethen and Embree [26].

During the early stages of the evolution, as the erf layer develops, the time step history in Figure 5.5 (top, □) is seen to follow that for the heat equation up until advection and diffusion have comparable magnitude. This occurs when the widths of the erf and exponential layers are comparable: $\sqrt{\nu t} \approx \nu/a$ which gives rise to what we refer to as the advection-diffusion time scale

$$\tau_{\mathrm{AD}} = \nu/a^2 \tag{5.8}$$

and its location is highlighted in Figure 5.5. This is also the time required for material to be transported through the width of the exponential boundary layer so is a measure of the time it takes to attain a steady state in the outflow layer. It is clearly a physical, rather than a numerical time scale—thus correcting the discussion in [7, section 2.6.2g]. The minimum time of believability on the fine grid can be expressed in terms of τ_{AD} and N as

$$\tau_{\mathrm{MTB}(h)} = \left(\frac{2\ln N}{N}\right)^2 \tau_{\mathrm{AD}}. \tag{5.9}$$

The difference $\|U - u_{\mathrm{erf}}\|_\infty$, shown in Figure 5.5 for a Shishkin grid (□), is a minimum when $t \approx 10^{-6}$. The behavior is almost identical for the geometric-Shishkin grid (▷).

The next time scale, which is a numerical artifact, occurs when the width of the erf layer ($\sqrt{\nu t}$) grows to that of the fine grid: $\sqrt{\nu t} = \beta$, leading to $\tau_2 = \beta^2/\nu$. There follows a period during which the numerical solution is stationary in the layer and a wave emanates to the left with speed $-3a$ and having the same oscillatory structure as that analyzed in section 5.1.

The progress of the solution to steady state is monitored in Figure 5.5 where we show $\|U - U_{\mathrm{SS}}\|_\infty$ as a function of time. It is small when $t \gtrapprox 10^{-3} = 10\tau_{\mathrm{AD}}$. The behavior is again almost identical for the geometric-Shishkin grid (▷). The nodal differences $U_j - u_{\mathrm{SS}}(x_j)$ between the numerical and exact solutions (5.7) at steady state are shown in Figure 5.6 (as a function of j) for Shishkin grid (dots) and geometric-Shishkin grid (solid line) with $N = 256$. It is seen that the difference is oscillatory for the Shishkin grid when $j \le N/2$ (outside the layer region)[12] but the dominant error occurs inside the layer region and is of almost identical magnitude for both grids.

When the geometric-Shishkin grid is employed the time step history (see Figure 5.5, top, ▷) is virtually identical to the classical Shishkin grid while the dynamics of the solution are restricted to the layer ($t < \tau_2$). Thereafter, the advection-diffusion solution evolves using a much larger time step with the geometric-Shishkin grid since the amplitude of the $2h$–wavelength wave propagating in the negative x-direction from the grid interface is much smaller when there is a smooth transition from fine to coarse grid.

[12] An explicit expression for the steady state solution on a Shishkin grid is readily obtained from which it can be shown that $|U_{N/2} - u_{\mathrm{SS}}(x_{N/2})| \approx 1/N^2$.

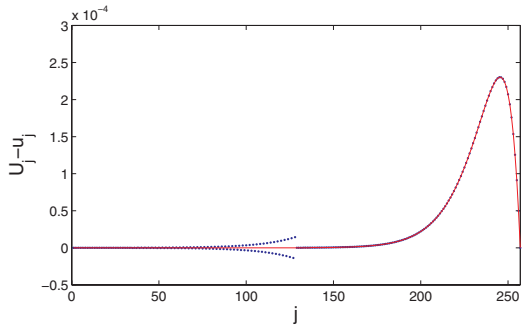

FIG. 5.6. *The difference between the nodal values of the steady state $u_{\rm SS}(x_j)$ of the advection-diffusion equation and the steady state nodal values U_j of the numerical solution on a Shishkin grid (dots) and geometric-Shishkin grid (solid line) for Example* 5.2, $N = 256$, $\nu = 10^{-4}$, *and* $a = 1$.

We complete this example by looking more closely at the transition from erf to exponential layer for a variety of physical parameters ν and a. For the continuum equation the early time is dominated by diffusion so $u_t \approx \nu u_{xx}$ (which leads to the approximate solution $u_{\rm erf}$) and, at long time, the advective and diffusive terms balance, i.e., $au_x \approx \nu u_{xx}$ (so $u_t \approx 0$, and this leads to the steady state solution $u_{\rm SS}$). It is convenient to rescale the time by $\tau = t/\tau_{\rm AD} \equiv a^2 t/\nu$ so that the advection-diffusion equation becomes

$$u_\tau + {\rm Pe}^{-1} u_x = {\rm Pe}^{-2} u_{xx}, \tag{5.10}$$

where ${\rm Pe} = a/\nu$ is the Peclet number. This nondimensionalization reduces the dependence of the solution to a single parameter. Since the advection speed is now ${\rm Pe}^{-1}$, the appropriate interval for time integration is $0 < \tau < O({\rm Pe})$. The FEM on a uniform grid will produce oscillation-free solutions provided that the grid Peclet number ${\rm Pe}_h = h{\rm Pe} = ah/\nu < 2$. A simulation is therefore said to be advection-dominated when ${\rm Pe}_h \gg 2$ (see, for instance, [7, pp. 216–217] for a discussion). On a Shishkin grid it is the coarse grid Peclet number, ${\rm Pe}_H$, that is relevant so advection dominates when ${\rm Pe} > N$.

In Figure 5.7 we show results for Peclet numbers ${\rm Pe} = 10^2$ ($\circ$), 10^3($\square$), and 10^4($*$). The time step $\Delta\tau = \Delta t/\tau_{\rm AD}$ increases slightly with Peclet number during the parabolic smoothing phase. The figure also shows $\|U - u_{\rm erf}\|_\infty$ as a function of $t/\tau_{\rm AD}$. It is seen that the value of the norm is essentially the same for all parameter values. We also show the results using $N = 512$ and ${\rm Pe} = 10^3$ (dashed curve, for which the minimum time of believability is $\tau = \tau_{{\rm MTB}(h)}/\tau_{\rm AD} \approx 5 \times 10^{-4}$, see (5.9)) and for $N = 256$ and ${\rm Pe} = 10$ (dotted curve) which suggest that the norm is reduced as N increases or Pe decreases up until $t \approx 5\tau_{{\rm MTB}(h)}$, where the norm attains its minimum. Thereafter the norm of the difference is independent of both N and Pe (provided advection dominates). The lower curve shows $\|U - U_{\rm SS}\|_\infty$ as a function of $t/\tau_{\rm AD}$ and the curves corresponding to the three Peclet numbers are again roughly coincident for all t (they are also independent of N and Pe for $t > \tau_{\rm MTB}(h)$ and ${\rm Pe} \gtrapprox 10$). The results in Figure 5.7 are consistent with the relationship (5.9) between time scales.

Figure 5.8 shows the erf solution (solid curve), steady state U_{SS} (dotted curve) and numerical solution U (dots and broken curve) in the four elements next to the outflow at ${\rm Pe} = 10^3$ for a range of times. Overall we observe that the differential equation is not accurately solved until $t \approx \tau_{{\rm MTB}(h)}$; the erf solution ($u_t \approx \nu u_{xx}$) then holds until t is a little beyond $0.01\tau_{\rm AD}$. All three terms in the differential equation

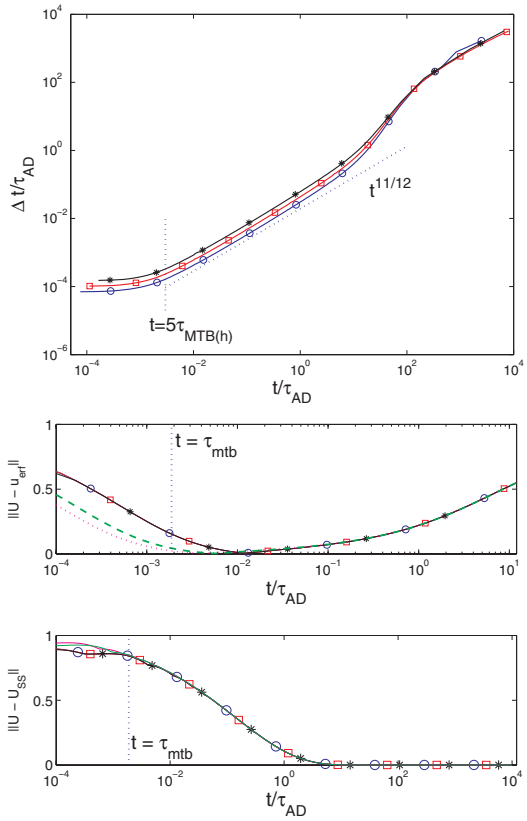

FIG. 5.7. *Advection-diffusion problem with step initial data on a geometric-Shishkin grid with* $N = 256$ *and* $\varepsilon = 10^{-7}$ *(Example* 5.2*) for* $\mathrm{Pe} = 10^2$ *(*$\circ$*),* $\mathrm{Pe} = 10^3$ *(*$\square$*),* $\mathrm{Pe} = 10^4$ *(*$*$*). Top:* $\Delta\tau$ *vs.* τ *for the rescaled equation* (5.10)*. Bottom:* $\|U - u_{\mathrm{erf}}\|_\infty$ *vs.* t/τ_{AD} *(dotted curve* $\mathrm{Pe} = 10$*, broken curve* $N = 512$*) and* $\|U - U_{\mathrm{SS}}\|_\infty$ *vs.* t/τ_{AD}*. The vertical dotted lines are drawn at* $t = \tau_{\mathrm{MTB}(h)}$*.*

have to be retained until steady state ($u_t \approx 0$) is approached at $\tau \approx 10$. There is no significant difference for any Peclet numbers $\mathrm{Pe} \geq 1$.

Our final example incorporates another combination of the time scales seen in Examples 5.1 and 5.2 but with more interesting "physics."

Example 5.3. Consider the system (1.4) arising from discretizing $u_t + au_x = \nu u_{xx}$ on $0 < x \leq 1$, with BCs $u(0,t) = 0$ and $u(1,t) = 0$ for $t \geq 0$, and the initial condition is the Gaussian profile (4.5) centered at the point $x = 1 - \sigma$ with $\sigma = 1/\sqrt{200}$ (see Figure 5.9).

The main differences between this and the previous example are that the solution is nonconstant in the outer region and the amplitude of the outflow boundary layer is also time varying.

With the Dirichlet outflow BC the effects of advection are again negligible at early times and the solution is given quite accurately by the approximation (cf. (3.15))

$$u(x,t) \approx u_e(x,t) \equiv u_0(x)\mathrm{erf}\big((1-x)/\sqrt{4\nu t}\big). \tag{5.11}$$

This erf layer then develops into an exponential layer at which stage the solution is given, again quite accurately, by the approximation

$$u(x,t) \approx u_\ell(x,t) \equiv u_{\mathrm{SS}}(x)u_\infty(x,t), \tag{5.12}$$

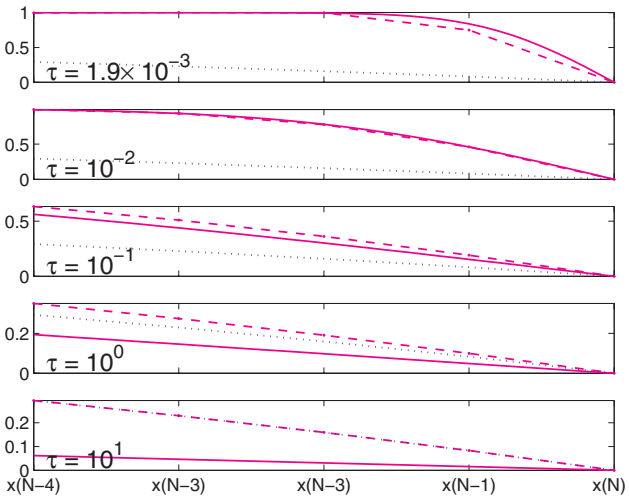

FIG. 5.8. *The erf solution (solid curve), steady state U_{SS} (dotted curve), and numerical solution U of* (5.10) *(dots and broken curve) in the four elements next to the outflow for Example* 5.2 *at times $\tau = t/\tau_{\mathrm{AD}} = 1.9 \times 10^{-3}(t = \tau_{\mathrm{MTB}(h)}), 10^{-2}, 10^{-1}, 1, 10$ with $N = 256$, $\varepsilon = 10^{-7}$, and $\mathrm{Pe} = 10^3$.*

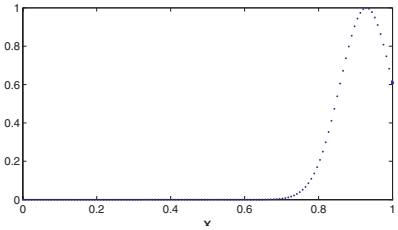

FIG. 5.9. *Initial condition for Example* 5.3.

where u_{SS} is given by (5.7) and u_∞ in footnote 11—the latter could equally be replaced by the exact solution for advection on an infinite span: $u_0(x - at)$. We shall be more precise about the time intervals over which these solutions are valid presently. The time-varying Gaussian (the "outer" solution) sets the amplitude of the solution in the boundary layer (the "inner" solution). It is noteworthy that, whereas u_∞ satisfies the full advection-diffusion equation and u_{SS} satisfies the steady state version of this equation, the solution given by (5.12) satisfies neither—yet does an excellent job of describing the physics, both within and outwith the boundary layer. A similar statement applies to (5.11).

The time step histories for the two Shishkin grids are plotted in Figure 5.10. The time step follows the familiar path through the fast transient $t \lessapprox \tau_{\mathrm{MTB}(h)}$ then increases as $t^{11/12}$ until $t \approx \tau_{\mathrm{AD}}$ after which it increases more rapidly as advection gains in strength. At $t = \tau_1$, Δt reaches the value given by (4.8)—advection is dominant and diffusive effects have little influence on its value. At this stage the outflow boundary layer is fully formed (i.e., is in steady state) with a slowly varying amplitude, but this variation has little effect on Δt since the width of the layer is so narrow that the solution within it makes a negligible contribution to the L_2 norm of $\ddot{U}$ for $\tau_1 < t < \tau_2$.

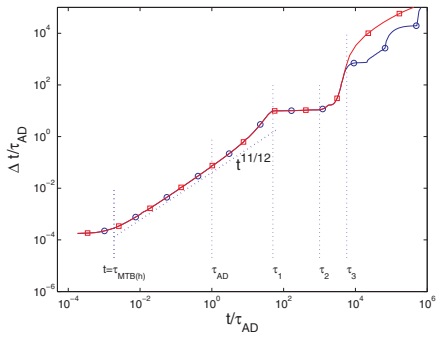

FIG. 5.10. *Time steps for Example* 5.3 *on a Shishkin grid* (∘) *and a geometric-Shishkin grid* (□) *with* $N = 256$, $\varepsilon = 10^{-7}$, $a = 1$, *and* $\nu = 10^{-4}$.

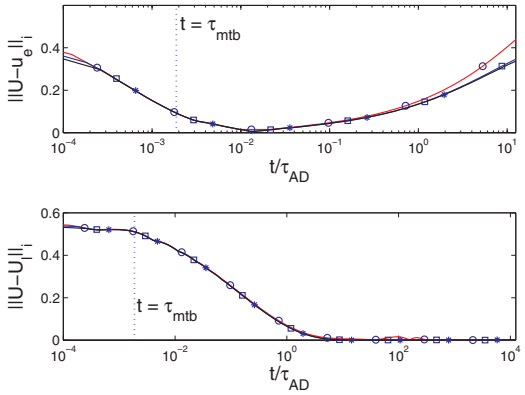

FIG. 5.11. *Top:* $\|U - u_{\mathrm{e}}\|_{\mathrm{i}}$, *Bottom:* $\|U - u_\ell\|$ *for the clipped-Gaussian initial data for both Shishkin and geometric-Shishkin grids with* $N = 256$, $\varepsilon = 10^{-7}$ *and* $\mathrm{Pe} = 10^3$ (∘), $\mathrm{Pe} = 10^4$ (□), *and* $\mathrm{Pe} = 10^5$ (∗).

For $t > \tau_2$ the increase in Δt becomes more rapid as the Gaussian "exits" the domain. Up to this time the two grids generate essentially identical histories. The constant τ_3 in Figure 5.10 is $O(\nu/a^2)$. Thus, for $t > \tau_3$, the numerical solutions become dominated by the spurious reflected waves from the grid interface and the time steps for the Shishkin grid, having much larger left-going waves, are appreciably smaller. Also noteworthy is the close similarity of the time step histories for Examples 5.2 and 5.3 (except of course for large t).

To estimate the times over which the two solutions, u_{e} and u_ℓ (given, respectively, by (5.11) and (5.12)) are valid we compute $\|U - u_{\mathrm{e}}\|_\infty$ and $\|U - u_\ell\|_\infty$ and these are shown in Figure 5.11 as functions of $\tau = t/\tau_{\mathrm{AD}}$ for $\mathrm{Pe} = 10^3$ (∘), $\mathrm{Pe} = 10^4$ (□), $\mathrm{Pe} = 10^5$ (∗). For fixed N both norms are essentially independent of Peclet number (in the advection dominated case). Both norms behave quantitatively as in Figure 5.7 when the solution in this example is scaled so that the initial amplitude of the discontinuity is unity—the same as for the step data. A detailed study of the solutions in the four elements closest to the outflow shows a behavior very similar to that in Figure 5.8 once the amplitude of the boundary layer solution is taken into account.

A corner singularity can also occur at $x = t = 0$ caused by the mismatch of the boundary data $u(0, t)$ and $u(x, 0)$ (or their derivatives) as $x, t \to 0$. The nature of the

singularity is discussed by Flyer and Fornberg [6] and the internal layer created as the effects are propagated into the domain along the characteristic $x = at$ is studied by Shih [25]. A finite element method with a fixed spatial grid is inappropriate in the case that a discontinuity occurs since this will generally create oscillatory numerical solutions. Weaker singularities on the other hand can be handled quite successfully and the behavior of the time step can be predicted from the level of regularity in the solution using the techniques described in section 3.

6. Possible extensions. Our examples reveal that even simple problems can have quite complex time scales, some physical and some of numerical origin, and in this paper we have endeavored, wherever possible, to identify as well as quantify the different phases of each simulation. It is clear that some form of adaptive time integrator is essential in order to efficiently respond to the different time scales and, given the wide range of dynamics taking place during these simulations, it is rather reassuring to see the TR-AB2 integrator find the appropriate time step during all phases. We have looked in detail at the way that smoothness of the initial data influences the solution, the error, and the selection of time steps. A close study of the behavior of the time step can often be useful in shedding light on the different temporal phases of a simulation.

We note that, of all A-stable linear multistep methods, TR has the smallest error constant and therefore allows the largest time step for a given accuracy. For the second order backward differentiation (BDF) BDF2 method (see Hairer, Norsett, and Wanner [12, p. 401] for the variable step formulation and Hundsdorfer and Verwer [15, p. 203] for numerical results) the error constant is $C_3 = -2/9$ from which we deduce that the time step selected by an adaptive time-stepping method will be $(3/8)^{1/3} \approx 0.72$ times smaller than that used by our TR-AB2 method. This has been verified by computation; for instance, in Example 3.2, the BDF2 time steps are smaller than those shown in Figure 3.3 by the predicted fraction up until $t \approx \tau_2$, after which both methods have approximately equal time steps (in keeping with our discussion of long term behavior in section 2). The same ratio of time steps is observed in pure advection problems provided that the tolerance is chosen to be sufficiently small that the time steps remain essentially constant; otherwise, the dissipative nature of BDF2 causes the time steps to increase with time.

The theoretical results of this paper have all been based on the principal truncation error term of the TR integrator. For a general pth order linear multistep method with error constant C_{p+1} (see, for instance, Hairer, Norsett, and Wanner [12]) we obtain, using (1.15),

$$\Delta t_n \approx \left(\frac{\varepsilon}{|C_{p+1}| \, \|\mathbf{u}^{(p+1)}\|} \right)^{\frac{1}{p+1}}.$$

Thus, for specific model problems, such as the examples used in this paper, it is possible to compare the efficiency of methods of differing orders as a prelude to the use of variable step-variable order methods. We intend to pursue these ideas in future publications.

Acknowledgments. Thanks are due to John Dold of the University of Manchester for improving our understanding of PDE asymptotics, and especially to Alan Hindmarsh who has been a guiding light to the "smart" time integration of ODE's (and DAE's, when Navier–Stokes is addressed) over the last thirty years; this paper is dedicated to him.

REFERENCES

[1] G. Akrivis, C. Makridakis, and R. Nochetto, *A posteriori error estimates for the Crank–Nicolson method for parabolic equations*, Math. Comp., 75 (2006), pp. 511–531.
[2] I. Babuška and T. Strouboulis, *The Finite Element Method and its Reliability*, Oxford University Press, New York, 2001.
[3] M. P. Calvo and J. M. Sanz-Serna, *Numerical Hamiltonian Problems*, Appl. Math. Math. Comput., 7, Chapman & Hall, 1994.
[4] T. F. Dupont, *A short survey of parabolic Galerkin methods*, in The Mathematical Basis of Finite Element Methods, Inst. Math. Appl. Conf. New Ser. 2, D. F. Griffiths, ed., University Press, Oxford, 1984, pp. 27–40.
[5] R. Fletcher and D. F. Griffiths, *The generalized eigenvalue problem for certain unsymmetric band matrices*, Linear Algebra Appl., 29 (1980), pp. 139–149.
[6] N. Flyer and B. Fornberg, *Accurate numerical resolution of transients in initial-boundary value problems for the heat equation*, J. Comput. Phys., 184 (2003), pp. 526–539.
[7] P. M. Gresho and R. L. Sani, *Incompressible Flow and the Finite Element Method, Vol.* 1: *Advection-Diffusion*, John Wiley & Sons, Chichester, UK, 2000.
[8] P. M. Gresho and R. L. Sani, *Incompressible Flow and the Finite Element Method, Vol.* 2: *Isothermal Laminar Flow*, John Wiley & Sons, Chichester, UK, 2000.
[9] D. F. Griffiths, *The dynamics of some linear multistep methods with step-size control*, in Numerical Analysis 1987, Pitman Res. Notes Math. Ser. 170, D. F. Griffiths and G. A. Watson, eds., Longman, Harlow, Sci. Tech., UK, 1988, pp. 115–134.
[10] D. F. Griffiths and J. M. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 994–1008.
[11] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations*, II, *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.
[12] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems*, Springer-Verlag, Berlin, 1993.
[13] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, New York, 1962.
[14] A. C. Hindmarsh, *private communication*, 2003.
[15] W. Hundsdorfer and J. G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin, 2003.
[16] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[17] H.-O. Kreiss and J. Oliger, *Methods for the approximate solution of time dependent problems*, GARP Publications Series 10, World Meteorological Organization, Geneva, 1973.
[18] M. Luskin and R. Rannacher, *On the smoothing property of the Crank–Nicolson scheme*, Applicable Anal., 14 (1982), pp. 117–135.
[19] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.
[20] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin, *Fitted Numerical Methods for Singular Perturbation Problems*, World Scientific, River Edge, NJ, 1996.
[21] K. W. Morton and D. F. Mayers, *Numerical Solution of Partial Differential Equations*, 2nd ed., Cambridge University Press, 2005.
[22] O. Osterby, *Five ways of reducing the Crank–Nicolson oscillations*, BIT, 43 (2003), pp. 811–822.
[23] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical methods for singularly perturbed differential equations*, in Convection Diffusion and Flow Problems, Springer Ser. Computat. Math. 24, Springer–Verlag, Berlin, 1996.
[24] L. F. Shampine, I. Gladwell, and S. Thompson, *Solving ODEs with MATLAB*, Cambridge University Press, Cambridge, UK, 2003.
[25] S.-D. Shih, *On a class of singularly perturbed parabolic equations*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 337–345.
[26] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra*, in The Behavior of Nonnormal Matrices and Operators, Princeton University Press, Princeton, NJ, 2005.
[27] R. Verfürth, *A posteriori error estimates for finite element discretizations of the heat equation*, Calcolo, 40 (2003), pp. 195–212.
[28] F. Verhulst, *Methods and Applications of Singular Perturbations: Boundary Layers and Multiple Timescale Dynamics*, Texts Appl. Math. 50, Springer, New York, 2005.
[29] R. F. Warming and B. J. Hyett, *The modified equation approach to the stability and accuracy analysis of finite-difference methods*, J. Comput. Phys., 14 (1974), pp. 159–179.